

(19) World Intellectual Property
Organization
International Bureau



(43) International Publication Date
1 April 2004 (01.04.2004)

PCT

(10) International Publication Number
WO 2004/027035 A2

- (51) International Patent Classification⁷: C12N (74) Agents: HODGES, Robert, A. et al.; Needle & Rosenberg, P.C., 999 Peachtree Street, Suite 1000, Atlanta, GA 33309-3915 (US).
- (21) International Application Number: PCT/US2003/029589
- (22) International Filing Date: 22 September 2003 (22.09.2003)
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data: 60/412,468 20 September 2002 (20.09.2002) US
- (71) Applicant: YALE UNIVERSITY [US/US]; 433 Temple Street, P.O.Box 208336, New Haven, CT 06520-8336 (US).
- (72) Inventors: BREAKER, Ronald, R.; 133 Weatherly Trail, Guilford, CT 06437 (US). NAHVI, Ali; 188 Lawrence Street #3, New Haven, CT 06511 (US). SUDARSAN, Narasimhan; 599A Prospect Street, New Haven, CT 06511 (US). EBERT, Margaret, S.; 107 West Broad Street, Hopewell, NJ 08525 (US). WINKLER, Wade; 433 Temple Street, P.O.Box 208336, New Haven, CT 06520-8336 (US). BARRICK, Jeffrey, E.; 169 Livingston Street, New Haven, CT 06511 (US). WICKISER, John, K.; 541 Prospect Street #3, New Haven, CT 06511 (US).
- (81) Designated States (*national*): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, UZ, VC, VN, YU, ZA, ZM, ZW.
- (84) Designated States (*regional*): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IT, LU, MC, NL, PT, RO, SE, SI, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).
- Published:**
— without international search report and to be republished upon receipt of that report
- For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.*

(54) Title: RIBOSWITCHES, METHODS FOR THEIR USE, AND COMPOSITIONS FOR USE WITH RIBOSWITCHES.

(57) Abstract: It has been discovered that certain natural mRNAs serve as metabolite-sensitive genetic switches wherein the RNA directly binds a small organic molecule. This binding process changes the conformation of the mRNA, which causes a change in gene expression by a variety of different mechanisms. Modified versions of these natural "riboswitches" (created by using various nucleic acid engineering strategies) can be employed as designer genetic switches that are controlled by specific effector compounds. Such effector compounds that activate a riboswitch are referred to herein as trigger molecules. The natural switches are targets for antibiotics and other small molecule therapies. In addition, the architecture of riboswitches allows actual pieces of the natural switches to be used to construct new nonimmunogenic genetic control elements, for example the aptamer (molecular recognition) domain can be swapped with other non-natural aptamers (or otherwise modified) such that the new recognition domain causes genetic modulation with user-defined effector compounds. The changed switches become part of a therapy regimen-turning on, or off, or regulating protein synthesis. Newly constructed genetic regulation networks can be applied in such areas as living biosensors, metabolic engineering of organisms, and in advanced forms of gene therapy treatments.

WO 2004/027035 A2

**RIBOSWITCHES, METHODS FOR THEIR USE,
AND COMPOSITIONS FOR USE WITH RIBOSWITCHES
CROSS-REFERENCE TO RELATED APPLICATIONS**

5 This application claims benefit of U.S. Provisional Application No. 60/412,468,
filed September 20, 2002. U.S. Provisional Application No. 60/412,468, filed September
20, 2002, is hereby incorporated herein by reference in its entirety.

STATEMENT REGARDING FEDERALLY SPONSORED RESEARCH

10 This invention was made with government support under Grants NIH GM48858
and NIH GM559343 awarded by the National Institutes of Health, and Grant NSF EIA-
0129939 awarded by the National Science Foundation. The government has certain
rights in the invention.

FIELD OF THE INVENTION

The disclosed invention is generally in the field of gene expression and
specifically in the area of regulation of gene expression.

15 **BACKGROUND OF THE INVENTION**

Precision genetic control is an essential feature of living systems, as cells must
respond to a multitude of biochemical signals and environmental cues by varying genetic
expression patterns. Most known mechanisms of genetic control involve the use of
protein factors that sense chemical or physical stimuli and then modulate gene
20 expression by selectively interacting with the relevant DNA or messenger RNA
sequence. Proteins can adopt complex shapes and carry out a variety of functions that
permit living systems to sense accurately their chemical and physical environments.
Protein factors that respond to metabolites typically act by binding DNA to modulate
transcription initiation (e.g. the lac repressor protein; Matthews, K.S., and Nichols, J.C.,
25 1998, *Prog. Nucleic Acids Res. Mol. Biol.* 58, 127-164) or by binding RNA to control
either transcription termination (e.g. the PyrR protein; Switzer, R.L., et al., 1999, *Prog.*
Nucleic Acids Res. Mol. Biol. 62, 329-367) or translation (e.g. the TRAP protein;
Babitzke, P., and Gollnick, P., 2001, *J. Bacteriol.* 183, 5795-5802). Protein factors
responds to environmental stimuli by various mechanisms such as allosteric modulation
30 or post-translational modification, and are adept at exploiting these mechanisms to serve
as highly responsive genetic switches (e.g. see Ptashne, M., and Gann, A. (2002). *Genes*
and Signals. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY).

In addition to the widespread participation of protein factors in genetic control, it is also known that RNA can take an active role in genetic regulation. Recent studies have begun to reveal the substantial role that small non-coding RNAs play in selectively targeting mRNAs for destruction, which results in down-regulation of gene expression (e.g. see Hannon, G.J. 2002, *Nature* 418, 244-251 and references therein). This process of RNA interference takes advantage of the ability of short RNAs to recognize the intended mRNA target selectively via Watson-Crick base complementation, after which the bound mRNAs are destroyed by the action of proteins. RNAs are ideal agents for molecular recognition in this system because it is far easier to generate new target-specific RNA factors through evolutionary processes than it would be to generate protein factors with novel but highly specific RNA binding sites.

Although proteins fulfill most requirements that biology has for enzyme, receptor and structural functions, RNA also can serve in these capacities. For example, RNA has sufficient structural plasticity to form numerous ribozyme domains (Cech & Golden, Building a catalytic active site using only RNA. In: *The RNA World* R. F. Gesteland, T. R. Cech, J. F. Atkins, eds., pp.321-350 (1998); Breaker, *In vitro* selection of catalytic polynucleotides. *Chem. Rev.* 97, 371-390 (1997)) and receptor domains (Osborne & Ellington, Nucleic acid selection and the challenge of combinatorial chemistry. *Chem. Rev.* 97, 349-370 (1997); Hermann & Patel, Adaptive recognition by nucleic acid aptamers. *Science* 287, 820-825 (2000)) that exhibit considerable enzymatic power and precise molecular recognition. Furthermore, these activities can be combined to create allosteric ribozymes (Soukup & Breaker, Engineering precision RNA molecular switches. *Proc. Natl. Acad. Sci. USA* 96, 3584-3589 (1999); Seetharaman et al., Immobilized riboswitches for the analysis of complex chemical and biological mixtures. *Nature Biotechnol.* 19, 336-341 (2001)) that are selectively modulated by effector molecules.

These properties of RNA are consistent with speculation (Gold et al., From oligonucleotide shapes to genomic SELEX: novel biological regulatory loops. *Proc. Natl. Acad. Sci. USA* 94, 59-64 (1997); Gold et al., SELEX and the evolution of genomes. *Curr. Opin. Gen. Dev.* 7, 848-851 (1997); Nou & Kadner, Adenosylcobalamin inhibits ribosome binding to *btuB* RNA. *Proc. Natl. Acad. Sci. USA* 97, 7190-7195 (2000); Gelfand et al., A conserved RNA structure element involved in the regulation of bacterial riboflavin synthesis genes. *Trends Gen.* 15, 439-442 (1999); Miranda-Rios et

al., A conserved RNA structure (*thi* box) is involved in regulation of thiamin biosynthetic gene expression in bacteria. *Proc. Natl. Acad. Sci. USA* **98**, 9736-9741 (2001); Stormo & Ji, Do mRNAs act as direct sensors of small molecules to control their expression? *Proc. Natl. Acad. Sci. USA* **98**, 9465-9467 (2001)) that certain mRNAs might employ allosteric mechanisms to provide genetic regulatory responses to the presence of specific metabolites. Although a thiamine pyrophosphate (TPP)-dependent sensor/regulatory protein had been proposed to participate in the control of thiamine biosynthetic genes (Webb & Downs, Characterization of *thiL*, encoding thiamin-monophosphate kinase, in *Salmonella typhimurium*. *J. Biol. Chem.* **272**, 15702-15707 (1997)), no such protein factor has been shown to exist.

Transcription of the *lysC* gene of *B. subtilis* is repressed by high concentrations of lysine (Kochhar, S., and Paulus, H. 1996, *Microbiol.* **142**:1635-1639; Mäder, U., et al., 2002, *J. Bacteriol.* **184**:4288-4295; Patte, J.C. 1996. Biosynthesis of lysine and threonine. In: *Escherichia coli and Salmonella: Cellular and Molecular Biology*, F.C. Neidhardt, et al., eds., Vol. 1, pp. 528-541. ASM Press, Washington, DC; Patte, J.-C., et al., 1998, *FEMS Microbiol. Lett.* **169**:165-170), but that no protein factor had been identified that served as the genetic regulator (Liao, H.-H., and Hseu, T.-H. 1998, *FEMS Microbiol. Lett.* **168**:31-36). The *lysC* gene encodes aspartokinase II, which catalyzes the first step in the metabolic pathway that converts L-aspartic acid into L-lysine (Belitsky, B.R. 2002. Biosynthesis of amino acids of the glutamate and aspartate families, alanine, and polyamines. In: *Bacillus subtilis and its Closest Relatives: from Genes to Cells*. A.L. Sonenshein, J.A. Hoch, and R. Losick, eds., ASM Press, Washington, D.C.).

BRIEF SUMMARY OF THE INVENTION

It has been discovered that certain natural mRNAs serve as metabolite-sensitive genetic switches wherein the RNA directly binds a small organic molecule. This binding process changes the conformation of the mRNA, which causes a change in gene expression by a variety of different mechanisms. Modified versions of these natural "riboswitches" (created by using various nucleic acid engineering strategies) can be employed as designer genetic switches that are controlled by specific effector compounds. Such effector compounds that activate a riboswitch are referred to herein as trigger molecules. The natural switches are targets for antibiotics and other small molecule therapies. In addition, the architecture of riboswitches allows actual pieces of the natural switches to be used to construct new non-immunogenic genetic control

elements, for example the aptamer (molecular recognition) domain can be swapped with other non-natural aptamers (or otherwise modified) such that the new recognition domain causes genetic modulation with user-defined effector compounds. The changed switches become part of a therapy regimen—turning on, or off, or regulating protein synthesis.

- 5 Newly constructed genetic regulation networks can be applied in such areas as living biosensors, metabolic engineering of organisms, and in advanced forms of gene therapy treatments.

Disclosed are isolated and recombinant riboswitches, recombinant constructs containing such riboswitches, heterologous sequences operably linked to such
10 riboswitches, and cells and transgenic organisms harboring such riboswitches, riboswitch recombinant constructs, and riboswitches operably linked to heterologous sequences. The heterologous sequences can be, for example, sequences encoding proteins or peptides of interest, including reporter proteins or peptides. Preferred riboswitches are, or are derived from, naturally occurring riboswitches.

- 15 Also disclosed are chimeric riboswitches containing heterologous aptamer domains and expression platform domains. That is, chimeric riboswitches are made up an aptamer domain from one source and an expression platform domain from another source. The heterologous sources can be from, for example, different specific riboswitches or different classes of riboswitches. The heterologous aptamers can also
20 come from non-riboswitch aptamers. The heterologous expression platform domains can also come from non-riboswitch sources.

- Also disclosed are compositions and methods for selecting and identifying compounds that can activate, deactivate or block a riboswitch. Activation of a riboswitch refers to the change in state of the riboswitch upon binding of a trigger
25 molecule. A riboswitch can be activated by compounds other than the trigger molecule and in ways other than binding of a trigger molecule. The term trigger molecule is used herein to refer to molecules and compounds that can activate a riboswitch. This includes the natural or normal trigger molecule for the riboswitch and other compounds that can activate the riboswitch. Natural or normal trigger molecules are the trigger molecule for
30 a given riboswitch in nature or, in the case of some non-natural riboswitches, the trigger molecule for which the riboswitch was designed or with which the riboswitch was selected (as in, for example, in vitro selection or in vitro evolution techniques). Non-natural trigger molecules can be referred to as non-natural trigger molecules.

Deactivation of a riboswitch refers to the change in state of the riboswitch when the trigger molecule is not bound. A riboswitch can be deactivated by binding of compounds other than the trigger molecule and in ways other than removal of the trigger molecule. Blocking of a riboswitch refers to a condition or state of the riboswitch where the presence of the trigger molecule does not activate the riboswitch.

Also disclosed are compounds, and compositions containing such compounds, that can activate, deactivate or block a riboswitch. Also disclosed are compositions and methods for activating, deactivating or blocking a riboswitch. Riboswitches function to control gene expression through the binding or removal of a trigger molecule.

Compounds can be used to activate, deactivate or block a riboswitch. The trigger molecule for a riboswitch (as well as other activating compounds) can be used to activate a riboswitch. Compounds other than the trigger molecule generally can be used to deactivate or block a riboswitch. Riboswitches can also be deactivated by, for example, removing trigger molecules from the presence of the riboswitch. A riboswitch can be blocked by, for example, binding of an analog of the trigger molecule that does not activate the riboswitch.

Also disclosed are compositions and methods for altering expression of an RNA molecule, or of a gene encoding an RNA molecule, where the RNA molecule includes a riboswitch, by bringing a compound into contact with the RNA molecule. Riboswitches function to control gene expression through the binding or removal of a trigger molecule. Thus, subjecting an RNA molecule of interest that includes a riboswitch to conditions that activate, deactivate or block the riboswitch can be used to alter expression of the RNA. Expression can be altered as a result of, for example, termination of transcription or blocking of ribosome binding to the RNA. Binding of a trigger molecule can, depending on the nature of the riboswitch, reduce or prevent expression of the RNA molecule or promote or increase expression of the RNA molecule.

Also disclosed are compositions and methods for regulating expression of an RNA molecule, or of a gene encoding an RNA molecule, by operably linking a riboswitch to the RNA molecule. A riboswitch can be operably linked to an RNA molecule in any suitable manner, including, for example, by physically joining the riboswitch to the RNA molecule or by engineering nucleic acid encoding the RNA molecule to include and encode the riboswitch such that the RNA produced from the engineered nucleic acid has the riboswitch operably linked to the RNA molecule.

Subjecting a riboswitch operably linked to an RNA molecule of interest to conditions that activate, deactivate or block the riboswitch can be used to alter expression of the RNA.

Also disclosed are compositions and methods for regulating expression of a naturally occurring gene or RNA that contains a riboswitch by activating, deactivating or blocking the riboswitch. If the gene is essential for survival of a cell or organism that harbors it, activating, deactivating or blocking the riboswitch can in death, stasis or debilitation of the cell or organism. For example, activating a naturally occurring riboswitch in a naturally occurring gene that is essential to survival of a microorganism can result in death of the microorganism (if activation of the riboswitch turns off or represses expression). This is one basis for the use of the disclosed compounds and methods for antimicrobial and antibiotic effects.

Also disclosed are compositions and methods for regulating expression of an isolated, engineered or recombinant gene or RNA that contains a riboswitch by activating, deactivating or blocking the riboswitch. The gene or RNA can be engineered or can be recombinant in any manner. For example, the riboswitch and coding region of the RNA can be heterologous, the riboswitch can be recombinant or chimeric, or both. If the gene encodes a desired expression product, activating or deactivating the riboswitch can be used to induce expression of the gene and thus result in production of the expression product. If the gene encodes an inducer or repressor of gene expression or of another cellular process, activation, deactivation or blocking of the riboswitch can result in induction, repression, or de-repression of other, regulated genes or cellular processes. Many such secondary regulatory effects are known and can be adapted for use with riboswitches. An advantage of riboswitches as the primary control for such regulation is that riboswitch trigger molecules can be small, non-antigenic molecules.

Also disclosed are compositions and methods for altering the regulation of a riboswitch by operably linking an aptamer domain to the expression platform domain of the riboswitch (which is a chimeric riboswitch). The aptamer domain can then mediate regulation of the riboswitch through the action of, for example, a trigger molecule for the aptamer domain. Aptamer domains can be operably linked to expression platform domains of riboswitches in any suitable manner, including, for example, by replacing the normal or natural aptamer domain of the riboswitch with the new aptamer domain. Generally, any compound or condition that can activate, deactivate or block the

riboswitch from which the aptamer domain is derived can be used to activate, deactivate or block the chimeric riboswitch.

Also disclosed are compositions and methods for inactivating a riboswitch by covalently altering the riboswitch (by, for example, crosslinking parts of the riboswitch or coupling a compound to the riboswitch). Inactivation of a riboswitch in this manner can result from, for example, an alteration that prevents the trigger molecule for the riboswitch from binding, that prevents the change in state of the riboswitch upon binding of the trigger molecule, or that prevents the expression platform domain of the riboswitch from affecting expression upon binding of the trigger molecule.

Also disclosed are methods of identifying compounds that activate, deactivate or block a riboswitch. For examples, compounds that activate a riboswitch can be identified by bringing into contact a test compound and a riboswitch and assessing activation of the riboswitch. If the riboswitch is activated, the test compound is identified as a compound that activates the riboswitch. Activation of a riboswitch can be assessed in any suitable manner. For example, the riboswitch can be linked to a reporter RNA and expression, expression level, or change in expression level of the reporter RNA can be measured in the presence and absence of the test compound. As another example, the riboswitch can include a conformation dependent label, the signal from which changes depending on the activation state of the riboswitch. Such a riboswitch preferably uses an aptamer domain from or derived from a naturally occurring riboswitch. As can be seen, assessment of activation of a riboswitch can be performed with the use of a control assay or measurement or without the use of a control assay or measurement. Methods for identifying compounds that deactivate a riboswitch can be performed in analogous ways.

Identification of compounds that block a riboswitch can be accomplished in any suitable manner. For example, an assay can be performed for assessing activation or deactivation of a riboswitch in the presence of a compound known to activate or deactivate the riboswitch and in the presence of a test compound. If activation or deactivation is not observed as would be observed in the absence of the test compound, then the test compound is identified as a compound that blocks activation or deactivation of the riboswitch.

Also disclosed are biosensor riboswitches. Biosensor riboswitches are engineered riboswitches that produce a detectable signal in the presence of their cognate

trigger molecule. Useful biosensor riboswitches can be triggered at or above threshold levels of the trigger molecules. Biosensor riboswitches can be designed for use in vivo or in vitro. For example, biosensor riboswitches operably linked to a reporter RNA that encodes a protein that serves as or is involved in producing a signal can be used in vivo
5 by engineering a cell or organism to harbor a nucleic acid construct encoding the riboswitch/reporter RNA. An example of a biosensor riboswitch for use in vitro is a riboswitch that includes a conformation dependent label, the signal from which changes depending on the activation state of the riboswitch. Such a biosensor riboswitch preferably uses an aptamer domain from or derived from a naturally occurring
10 riboswitch. Also disclosed are methods of detecting compounds using biosensor riboswitches. The method can include bringing into contact a test sample and a biosensor riboswitch and assessing the activation of the biosensor riboswitch. Activation of the biosensor riboswitch indicates the presence of the trigger molecule for the biosensor riboswitch in the test sample.

15 Also disclosed are compounds made by identifying a compound that activates; deactivates or blocks a riboswitch and manufacturing the identified compound. This can be accomplished by, for example, combining compound identification methods as disclosed elsewhere herein with methods for manufacturing the identified compounds. For example, compounds can be made by bringing into contact a test compound and a
20 riboswitch, assessing activation of the riboswitch, and, if the riboswitch is activated by the test compound, manufacturing the test compound that activates the riboswitch as the compound.

Also disclosed are compounds made by checking activation, deactivation or blocking of a riboswitch by a compound and manufacturing the checked compound.
25 This can be accomplished by, for example, combining compound activation, deactivation or blocking assessment methods as disclosed elsewhere herein with methods for manufacturing the checked compounds. For example, compounds can be made by bringing into contact a test compound and a riboswitch, assessing activation of the riboswitch, and, if the riboswitch is activated by the test compound, manufacturing the
30 test compound that activates the riboswitch as the compound. Checking compounds for their ability to activate, deactivate or block a riboswitch refers to both identification of compounds previously unknown to activate, deactivate or block a riboswitch and to

assessing the ability of a compound to activate, deactivate or block a riboswitch where the compound was already known to activate, deactivate or block the riboswitch.

Also disclosed are methods for selecting, designing or deriving new riboswitches and/or new aptamers that recognize new trigger molecules. Such methods can involve
5 production of a set of aptamer variants in a riboswitch, assessing the activation of the variant riboswitches in the presence of a compound of interest, selecting variant riboswitches that were activated (or, for example, the riboswitches that were the most highly or the most selectively activated), and repeating these steps until a variant
10 riboswitch of a desired activity, specificity, combination of activity and specificity, or other combination of properties results. Also disclosed are riboswitches and aptamer domains produced by these methods.

The disclosed riboswitches, including the derivatives and recombinant forms thereof, generally can be from any source, including naturally occurring riboswitches and riboswitches designed de novo. Any such riboswitches can be used in or with the
15 disclosed methods. However, different types of riboswitches can be defined and some such sub-types can be useful in or with particular methods (generally as described elsewhere herein). Types of riboswitches include, for example, naturally occurring riboswitches, derivatives and modified forms of naturally occurring riboswitches, chimeric riboswitches, and recombinant riboswitches. A naturally occurring riboswitch
20 is a riboswitch having the sequence of a riboswitch as found in nature. Such a naturally occurring riboswitch can be an isolated or recombinant form of the naturally occurring riboswitch as it occurs in nature. That is, the riboswitch has the same primary structure but has been isolated or engineered in a new genetic or nucleic acid context. Chimeric riboswitches can be made up of, for example, part of a riboswitch of any or of a
25 particular class or type of riboswitch and part of a different riboswitch of the same or of any different class or type of riboswitch; part of a riboswitch of any or of a particular class or type of riboswitch and any non-riboswitch sequence or component. Recombinant riboswitches are riboswitches that have been isolated or engineered in a new genetic or nucleic acid context.

30 Different classes of riboswitches refer to riboswitches that have the same or similar trigger molecules or riboswitches that have the same or similar overall structure (predicted, determined, or a combination). Riboswitches of the same class generally, but

need not, have both the same or similar trigger molecules and the same or similar overall structure.

Additional advantages of the disclosed method and compositions will be set forth in part in the description which follows, and in part will be understood from the description, or can be learned by practice of the disclosed method and compositions. The advantages of the disclosed method and compositions will be realized and attained by means of the elements and combinations particularly pointed out in the appended claims. It is to be understood that both the foregoing general description and the following detailed description are exemplary and explanatory only and are not restrictive of the invention as claimed.

BRIEF DESCRIPTION OF THE DRAWINGS

The accompanying drawings, which are incorporated in and constitute a part of this specification, illustrate several embodiments of the disclosed method and compositions and together with the description, serve to explain the principles of the disclosed method and compositions.

Figures 1A and 1B show metabolite-dependent conformational changes in the 202-nucleotide leader sequence of the *btuB* mRNA. Figure 1A shows separation of spontaneous RNA-cleavage products of the *btuB* leader using denaturing 10% polyacrylamide gel electrophoresis (PAGE). 5'-32p-labeled mRNA leader molecules (arrow) were incubated for 41 hr at 25°C in 20 mM MgCl₂, 50 mM Tris-HCl (pH 8.3 at 25°C) in the presence (+) or absence (-) of 20 μM of AdoCbl. Lanes containing RNAs that have undergone no reaction, partial digest with alkali, and partial digest with RNase T1 (G-specific cleavage) are identified by NR, OH, and T1, respectively. The location of product bands corresponding to cleavage after selected guanosine residues are identified by filled arrowheads. Light blue arrowheads labeled 1 through 8 identify eight of the nine locations that exhibit effector-induced structure modulation, which experience an increase or decrease in the rate of spontaneous RNA cleavage. The image was generated using a phosphorimager (Molecular Dynamics), and cleavage yields were quantitated by using ImageQuant software. Figure 1B shows sequence and secondary-structure model for the 202-nucleotide leader sequence of *btuB* mRNA in the presence of AdoCbl. Putative base-paired elements are designated P1 through P9. Complementary nucleotides in the loops of P4 and P9 that have the potential to form a pseudoknot are juxtaposed. Nine specific sites of structure modulation are identified by light blue arrowheads. The

asterisks demark the boundaries of the B₁₂ box (nucleotides 141-162). The coding region and the 38 nucleotides that reside immediately 5' of the start codon (nucleotides 241-243) were not included in the 202-nucleotide fragment. The 315-nucleotide fragment includes the 202-nucleotide fragment, the remaining 38 nucleotides of the leader sequence, and the first 75 nucleotides of the coding region.

Figures 2A and 2B show the *btuB* mRNA leader forms a saturable binding site for AdoCbl. Figure 2A shows the dependence of spontaneous cleavage of *btuB* mRNA leader on the concentration of AdoCbl effector as represented by site 1 (G23) and site 2 (U68). 5'-³²P-labeled mRNA leader molecules were incubated, separated, and analyzed as described in the in the brief description of Figure 1, and include identical control and marker lanes as indicated. Incubations contained concentrations of AdoCbl ranging from 10 nM to 100 μ M (lanes 1 through 8) or did not include AdoCbl (-). Figure 2B shows a composite plot of the fraction of RNA cleaved at six locations along the mRNA leader versus the logarithm of the concentration (c) of AdoCbl. Fraction cleaved values were normalized relative to the highest and lowest cleavage values measured for each location, including the values obtained upon incubation in the absence of AdoCbl. The inset defines the symbols used for each of six sites, while the remaining three sites were excluded from the analysis due to weak or obscured cleavage bands. Filled and open symbols represent increasing and decreasing cleavage yields, respectively, upon increasing the concentration of AdoCbl. The dashed line reflects a K_D of ~300 nM, as predicted by the concentration needed to generate half-maximal structural modulation. Data plotted were derived from a single PAGE analysis, of which two representative sections are depicted in Figure 1A.

Figure 3 shows the 202-nucleotide mRNA leader causes an unequal distribution of AdoCbl in an equilibrium dialysis apparatus. I: Equilibration of tritiated effector was conducted in the absence of RNA. II: (step 1) Equilibration was conducted as in I, but with 200 pmoles of mRNA leader added to chamber b; (step 2) 5,000 pmoles of unlabeled AdoCbl was added to chamber b. III: Equilibrations were conducted as described in II, but wherein 5,000 pmoles of cyanocobalamin was added to chamber b. IV: (step 1) Equilibration was initiated as described in step 1 of II; (steps 2 and 3) the solution in chamber a was replaced with 25 μ L of fresh equilibration buffer; (step 4) 5,000 pmoles of unlabeled AdoCbl was added to chamber b. The cpm ratio is the ratio of

counts detected in chamber b relative to that of a. The dashed line represents a cpm ratio of 1, which is expected if equal distribution of tritium is established.

Figures 4A and 4B show selective molecular recognition of effectors by the *btuB* mRNA leader. Figure 4A shows a chemical structure of AdoCbl (1) and various effector analogs (2 through 11, ref. 30). Figure 4B shows a determination of analog binding by monitoring modulation of spontaneous cleavage of the 202-nucleotide *btuB* RNA leader. 5'-³²P-labeled mRNA leader molecules were incubated, separated, and analyzed as described in the legend to Figure 1A, and include identical control and marker lanes as indicated. The sections of three PAGE analyses encompassing site 2 (U68) are depicted. Below each image is plotted the amount of RNA cleaved (normalized with relation to the lowest and highest levels of cleavage at U68 in each gel) for each effector as indicated, or for no effector (-). The compound 11 (13-epi-AdoCbl) is an epimer of AdoCbl wherein the configuration at C13 is inverted, so that the e propionamide side chain is above the plane of the corrin ring; see Brown et al., Conformational studies of 5'-deoxyadenosyl-13-epicobalamin, a coenzymatically active structural analog of coenzyme B₁₂. Polyhedron 17, 2213 (1998).

Figures 5A, 5B, 5C, 5D, 5E and 5F show mutations in the mRNA leader and their effects on AdoCbl binding and genetic control. Figure 5A shows sequence of the putative P5 element of the wild-type 202-nucleotide *btuB* leader exhibits AdoCbl-dependent modulation of structure as indicated by the observed increase in spontaneous RNA cleavage at position U68 (10% denaturing PAGE gel). Assays were conducted in the absence (-) or presence (+) of 5 μ M AdoCbl. The remaining lanes are as described in the legend to Figure 1A. The composite bar graph reflects the ability of the RNA to shift the equilibrium of AdoCbl in an equilibrium dialysis apparatus and the ability of a reporter gene (see Experimental Procedures) to be regulated by AdoCbl addition to a bacterial culture. (Left) Plotted is the cpm ratio derived by equilibrium dialysis, wherein chamber b contains the RNA. Details of the equilibrium dialysis experiments are described in the brief description of Figure 3. (Right) Plotted are the expression levels of β -galactosidase as determined from cells grown in the absence (-) or presence (+) of 5 μ M AdoCbl. Boxed numbers on the left and right, respectively, reflect the approximate K_D and the fold repression of β -galactosidase activity in the presence of AdoCbl. N.D. designates not determined. Figure 5B-5F shows sequences and performance

characteristics of various mutant leader sequences as indicated. Constructs were created as described in the Experimental Procedures section.

Figures 6A, 6B, 6C and 6D show metabolite binding by mRNAs. Figure 6A shows TPP-dependent modulation of the spontaneous cleavage of 165 *thiM* RNA was visualized by polyacrylamide gel electrophoresis (PAGE). 5' ³²P-labeled RNAs (arrow, 20 nM) were incubated for approximately 40 hr at 25°C in 20 mM MgCl₂, 50 mM Tris-HCl (pH 8.3 at 25°C) in the presence (+) or absence (-) of 100 μM TPP. NR, ⁻OH and T1 represent RNAs subjected to no reaction, partial digestion with alkali, or partial digestion with RNase T1 (G-specific cleavage), respectively. Product bands representing cleavage after selected G residues are numbered and identified by filled arrowheads. The asterisk identifies modulation of RNA structure involving the Shine-Dalgarno (SD) sequence. Gel separations were analyzed using a phosphorimager (Molecular Dynamics) and quantitated using ImageQuant software. Figure 6B shows a secondary-structure model of 165 *thiM* as predicted by computer modeling (Zuker et al., Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide. In *RNA Biochemistry and Biotechnology* (eds. Barciszewski J. & Clark, B.F.C.) 11-43 (NATO ASI Series, Kluwer Academic Publishers, 1999); Mathews et al., Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* 288, 911-940 (1999)) and by the structure probing data depicted in Figure 6A. Spontaneous cleavage characteristics are as noted in the inset. Unmarked nucleotides exhibit a constant but low level of degradation. The truncated 91 *thiM* RNA is boxed and the *thi* box element (Miranda-Rios et al., A conserved RNA structure (*thi* box) is involved in regulation of thiamin biosynthetic gene expression in bacteria. *Proc. Natl. Acad. Sci. USA* 98, 9736-9741 (2001)) is shaded light blue. Nucleotides highlighted in orange identify an alternative pairing, designated P8*. The RNA carries two mutations (G156A and U157C) relative to wild type that were introduced in a non-essential portion of the construct to form a restriction site for cloning, while all RNAs carry two 5'-terminal G residues to facilitate *in vitro* transcription. Figure 6C shows TPP-dependent modulation of the spontaneous cleavage of 240 *thiC* RNA. Reactions were conducted and analyzed as described in above for Figure 6A. Figure 6D shows a secondary-structure model of 240 *thiC*. Base-paired elements that are similar to those of *thiM* are labeled P1 through P5. The truncated RNA 111 *thiC* is boxed. Nucleotides highlighted in orange identify an alternative pairing.

Figures 7A, 7B and 7C show the *thiM* and *thiC* mRNA leaders serve as high-affinity metabolite receptors. Figure 7A shows the extent of spontaneous modulation of RNA cleavage at several sites within 165 *thiM* (left) and 240 *thiC* (right) plotted for different concentrations (*c*) of TPP. Red arrows reflect the estimated concentration of TPP needed to attain half maximal modulation of RNA (apparent K_D). Figure 7B shows the logarithm of the apparent K_D values plotted for both RNAs with TPP, TP and thiamine as indicated. The boxed data was generated using TPP with the truncated RNAs 91 *thiM* and 111 *thiC*. Figure 7C shows that patterns of spontaneous cleavage of 165 *thiM* differ between thiamine and TPP ligands as depicted by PAGE analysis (left) and as reflected by graphs (right) representing the relative phosphorimager counts for the three lanes as indicated. Details for the RNA probing analysis are similar to those described above in connection with Figure 6A. The graphs were generated by ImageQuant software.

Figures 8A, 8B and 8C show high sensitivity and selectivity of mRNA leaders for metabolite binding. Figure 8A shows chemical structures of several analogues of thiamine. TD is thiamine disulfide and THZ is 4-methyl-5- β -hydroxyethylthiazole. Figure 8B shows PAGE analysis of 165 *thiM* RNA structure probing using TPP and various chemical analogues (40 μ M each) as indicated. Locations of significant structural modulation within the RNA spanning nucleotides ~113 to ~150 are indicated by open arrowheads. The asterisk identifies the site (C144) used to compare the normalized fraction of RNA that is cleaved (bottom) in the presence of specific compounds. Details for the RNA probing analysis are similar to those described above in connection with Figure 6a. Figure 8C shows a summary of the features of TPP that are critical for molecular recognition. Figure 8D shows equilibrium dialysis using 3 H-thiamine as a tracer. Plotted are the ratios for tritium distribution in a two-chamber system (*a* and *b*) that were established upon equilibration in the presence of the RNA constructs in chamber *b* as indicated (see below for a description of the non-TPP-binding mutant M3). 100 μ M TPP or oxythiamine were added to chamber *a*, as denoted, upon the start of equilibration.

Figures 9A, 9B and 9C show mutational analysis of the structure and function of the *thiM* riboswitch. Figure 9A shows mutations present in constructs M1 through M8 relative to the 165 *thiM* RNA. P8* is a putative base-paired element between portions

(orange) of the P1 and P8 stems. Figure 9B (top) shows *in vitro* ligand-binding and genetic control functions of the wild-type (WT), M1 and M2 RNAs as reflected by PAGE analysis of in-line probing experiments (10 μ M TPP) and by β -galactosidase expression assays. Labels on PAGE gels are as described above in connection with Figure 6A. Bars represent the levels of gene expression in the presence (+) and the absence (-) of TPP in the culture medium. Figure 9C is a summary of similar analyses of WT through M9 is presented in table form. The SD status "n.d." (not determined) indicates either that the level of spontaneous cleavage detected in the absence and presence of TPP is near the limit of detection (M6, M7 and M8) or that the region adopts an atypical structure (M9) compared to WT.

Figure 10 shows a construct for the selection of SAM-responsive ribozymes. The hammerhead self-cleaving ribozyme and the SAM aptamer both require proper formation of the bridge domain to exhibit function. Therefore, the selection is expected to permit ribozyme function only when SAM or another binding-competent analog is present.

Figures 11A, 11B, 11C, 11D, 11E, 11F and 11G show consensus sequences and putative secondary structures were derived by phylogenetic and biochemical analyses as described for each riboswitch (see references). Nucleotides in red are conserved in greater than 90% of the representative sequences, open circles identify nucleotide positions of variable sequence, and lines identify elements that are variable in sequence and length. Models are described as follows: 6A) coenzyme B12 aptamer (Example 1); 6B) TPP aptamer (Example 2); 6C) FMN aptamer (Example 3); 6D) SAM aptamer (Example 7); 6E) guanine aptamer (Example 6); 6F) adenine aptamer (Example 8); and 6G) lysine aptamer (Example 5). Letters R and Y represent purine and pyrimidine bases, respectively; K designates G or U; W designates A or U; H designates A, C, or U; D designates G, A, or U; N represents any of the four bases.

Figures 12A, 12B and 12C show the regulation of the *B. subtilis* ribD mRNA by FMN. Figure 12A shows the results of in-line probing assays. Internucleotide linkages identified with red circles exhibit decreased amounts of spontaneous cleavage when ribD is incubated in the presence of FMN (indicating an increase in order for these nucleotides) relative to incubation in the absence of FMN. Yellow circles identify linkages that exhibit consistently high levels of scission, which indicates they are not modulated by presence of FMN. Figure 12B shows a model for the mechanism of ribD

regulation. The ribD mRNA adopts anti-termination conformation in the absence of FMN. Increased levels of FMN stabilize an RFN-FMN complex that permits formation of the terminator structure. Figure 12C shows the chemical structure and apparent dissociation constants for riboflavin and FMN.

5 Figures 13A, 13B and 13C show the regulation of the *E. coli* thiM mRNA by TPP. Figure 13A shows results of in-line probing assays. Internucleotide linkages identified with red circles exhibit decreased amounts of spontaneous cleavage when thiM is incubated in the presence of TPP compared to incubation in the absence of ligand. In contrast, linkages identified with green circles exhibit increased amounts of cleavage
10 when thiM is incubated with TPP compared to incubation in the absence of ligand. Blue-shaded box indicates pyrophosphate-recognition region (as described in text). Figure 13B shows a model for the mechanism of thiM regulation. In the absence of TPP, the anti-SD sequence interacts with part of aptamer domain to form anti-anti-SD. As TPP is increased, aptamer-TPP complexes are formed and the anti-SD favors pairing with the
15 SD. Figure 13C shows the chemical structure and apparent dissociation constants for thiamine and TPP.

 Figures 14A, 14B and 14C show putative eukaryote riboswitches. Figure 14A shows the consensus TPP binding domain based on 100 bacteria and archaea RNAs. Nucleotides in red are most conserved (>90%). Open circles represent nucleotide
20 positions and domains that vary in sequence and length are designated *var*. The consensus model is similar to that reported recently (Rodionov et al., 2002). Figure 14B shows the TPP-binding domain of *A. thaliana*. Variations in *O. sativa* (orange) and *P. secunda* (green) are shown. Figure 14C shows a putative TPP-binding domain in the intron of *N. crassa*.

25 Figure 15 shows sequence alignments of eukaryotic domains related to bacterial TPP-dependent riboswitches. Base paired stems are shaded in black and labeled as defined in Example 2). The P3 sequences, which in eukaryotes are significantly expanded in length and number of base pairs, are represented as a stem-loop structure. The highly conserved nucleotide positions in bacteria that were used to search for
30 eukaryotic domains are shaded gray. For each identified (ID) sequence, the position of the conserved CUGAGA sequence within the given Genbank entry is given along with the accession identification, sequence name, and gene identification. Additional protein annotations based on sequence similarity are shown in brackets. Methods: Riboswitch-

like domains were initially identified by sequence similarity to bacterial sequences (Eco2 and Cac) by a blastn search of Genbank using default parameters. These hits were verified and expanded by searching for degenerate matches to the pattern (CTGAGA [200] ACYTGA [5] <<< GNTNNNNC >>> [5] CGNRGGRA). Angle brackets indicate base pairing and bracketed numbers are variable gaps with constrained maximum lengths. All of the eukaryotic sequences have one or zero mismatches to this pattern except for one (Aor) that initially had three mismatches due to a single A insertion in the final search element. This mutation was removed to simplify the alignment. Comparison of mRNA (M33643.1) and genomic (AB033416.1) sequences demonstrated that the *F. oxysporum* element is in an intron in the 5' UTR of the sti35 gene. Other fungal sequences (Ncr, Aor, and Fso) are flanked by consensus splicing sequences.

Figures 16A and 16B show the structural probing of the putative TPP-riboswitch from *Arabidopsis*. Figure 16A shows the fragmentation pattern of the 128-nucleotide RNA (arrow) of *A. thaliana* (Fig. 14B) which was generated by incubation in the absence (-) or presence (+) of 100 μ M TPP. T1, $\bar{\text{OH}}$ and NR identify RNAs that were partially digested with RNase T1 (cleaves 3' to G residues), alkali, or were not reacted, respectively. Reactions were conducted as described in Example 2. Figure 16B shows the apparent K_D for TPP binding by the *A. thaliana* RNA. Fraction bound was determined by in-line probing as described in Examples 1-3.

Figure 17 shows genetic structures thiamine biosynthetic genes and possible mechanisms of riboswitch control. The location and mechanism of the *E. coli* and *B. subtilis* riboswitches are detailed in Examples 2 and 6. The putative TPP riboswitch from *P. secunda* resides immediately upstream from the polyA tail in the cDNA clone of the THIC gene. The putative TPP riboswitch domain in *F. oxysporum* is located in a 5'-UTR intron of the STI35 gene according to the genomic sequence but is absent in the cDNA clone.

Figure 18 shows the L box - a highly conserved sequence and structural domain is present in the 5'-UTRs of Gram-positive and Gram-negative bacterial mRNAs that are related to lysine metabolism. Conserved portions of the L box sequence and secondary structure were identified by alignment of 28 representative mRNAs as noted. Base pairing potential representing P1 through P5 are individually colored. Nucleotides in red are conserved in greater than 80% of the examples. The asterisk identifies the representative (*B. subtilis* *lysC* 5'-UTR) that was examined in this study. Gene names are

as annotated in GenBank or were derived by protein sequence similarity. Organism abbreviations are as follows: *Bacillus anthracis* (BA), *Bacillus halodurans* (BH), *Bacillus subtilis* (BS), *Clostridium acetobutylicum* (CA), *Clostridium perfringens* (CP), *Escherichia coli* (EC), *Haemophilus influenzae* (HI), *Oceanobacillus iheyensis* (OI),
 5 *Pasteurella multocida* (PM), *Staphylococcus aureus* (SA), *Staphylococcus epidermidis* (SE), *Shigella flexneri* (SF), *Shewanella oneidensis* (SO), *Thermatoga maritima* (TM), *Thermoanaerobacter tengcongensis* (TT), *Vibrio cholerae* (VC), *Vibrio vulnificus* (VV), *Thermoanaerobacter tengcongensis* (TE).

Figures 19A, 19B and 19C show the consensus L box motif from the *lysC* 5'-
 10 UTR of *B. subtilis* undergoes allosteric rearrangement in the presence of L-lysine. (A) Consensus sequence and structure of the L box domain as derived using a phylogeny of 31 representative sequences from prokaryotic and archaeal organisms (Fig. 18). Nucleotides depicted in red are present in at least 80% of the representatives, open circles identify nucleotide positions of variable identity, and tan lines denote variable
 15 nucleotide identity and chain length. Figure 19B shows sequence, secondary structure model, and lysine-induced structural modulation of the *lysC* 5'-UTR of *B. subtilis*. An additional 94 nucleotides (not depicted) reside between nucleotide 237 and the AUG start codon. Structural modulation sites (red-encircled nucleotides) were established using 237 *lysC* RNA by monitoring spontaneous RNA cleavage as depicted in C. Figure
 20 19C shows in-line probing of the 237 *lysC* RNA reveals lysine-induced modulation of RNA structure. Patterns of spontaneous cleavage, revealed by product separation using denaturing 10% polyacrylamide gel electrophoresis (PAGE), are altered at four major sites (denoted 1 through 4) in the presence (+) of 10 μ M L-lysine (L) relative to that observed in the absence (-) of lysine. T1, OH and NR represent partial digest with
 25 RNase T1, partial digest with alkali, and no reaction, respectively. Selected bands in the T1 lane (G-specific cleavage) are identified by nucleotide position. See Methods for experimental details.

Figures 20A, 20B, 20C, 20D and 20E show the molecular recognition characteristics of the lysine aptamer and the use of caged lysine. Figure 20A shows the
 30 chemical structures of L-lysine, D-lysine and nine closely-related analogs. Small circles represent chiral carbon centers wherein the enantiomeric configuration is defined for each compound. Shaded atoms identify chemical differences between L-lysine and the analog depicted. Figure 20B shows in-line probing analysis of the 179 *lysC* RNA in the

absence (–) of ligand, or in the presence of 10 μ M L-lysine or 100 μ M of various analogs as indicated for each lane. For each lane, the relative extent of spontaneous cleavage at site 3 is compared to that of the zone of constant cleavage immediately below this site, where a cleavage ratio significantly below ~1.5 reflects modulation. Figure 20C shows a schematic representation of dipeptide digestion by hydrochloric acid. All dipeptide forms are expected to be incapable of binding the lysine aptamer (*inactive*), while lysine-containing dipeptides should induce conformational changes in the aptamer (*active*) upon acid digestion. Figure 20D shows in-line probing analysis of the 179 *lysC* RNA in the absence of lysine (–) or in the presence of various amino acids and dipeptides.

Underlined lanes carry dipeptide preparations that were pretreated with HCl as depicted in a. Figure 20E shows the fraction of spontaneous cleavage at site 3 in d is plotted after normalization to the extent of processing in the absence of added ligand.

Figures 21A, 21B, 21C and 21D show determination of the dissociation constant and stoichiometry for L-lysine binding to the 179 *lysC* RNA. Figure 21A shows in-line probing with increasing concentrations of L-lysine ranging from 3 nM to 3 mM. Details are as defined for Fig. 19C. Figure 20B shows a plot depicting the normalized fraction of RNA undergoing spontaneous cleavage versus the concentration of amino acid for sites 1 through 3. The dashed line identifies the concentration of L-lysine required to bring about half-maximal structural modulation, which indicates the apparent K_D for ligand binding. Figure 20C shows the 179 *lysC* RNA (10 μ M) shifts the equilibrium of tritiated L-lysine (50 nM) in an equilibrium dialysis chamber. To investigate competitive binding, unlabeled L- (L) and D-lysine (D), or L-ornithine (O) were added to a final concentration of 50 μ M each to one chamber of a pre-equilibrated assay as indicated. Figure 21D shows a scatchard analysis of L-lysine binding by the 179 *lysC* RNA. The variable r represents the ratio of bound ligand concentration versus the total RNA concentration and the variable $[L_F]$ represents the concentration of free ligand.

Figures 22A, 22B and 22C show the *B. subtilis lysC* riboswitch and its mechanism for metabolite-induced transcription termination. Figure 22A shows a sequence and repressed-state model for the *lysC* riboswitch secondary structure. The nucleotides highlighted in orange identify the putative anti-terminator interaction that could form in the absence of L-lysine. Boxed nucleotides identify sites of disruption (M1) and compensatory mutations for the terminator stem (M2) and for the terminator and anti-terminator stems (M3). Nucleotides shaded in light blue identify some of the

positions where mutations exhibit *lysC* derepression that were reported previously (Vold et al. 1975; Lu et al. 1992). Figure 22B shows *In vitro* transcription assays conducted in the absence (–) or presence (+) of 10 mM L-lysine or other analogs as indicated. FL and T identify the full-length and terminated transcripts, respectively. The percent of the terminated RNAs relative to the total terminated and full-length transcripts are provided for each lane (% term.). Figure 22C shows *In vivo* expression of a β -galactosidase reporter gene fused to wild-type (WT), G39A and G40A mutant *lysC* 5'-UTR fragments. Media conditions are as follows: I, normal medium (0.27 mM lysine); II, minimal medium (0.012 mM); III, lysine-supplemented minimal medium (1 mM); IV, lysine hydroxamate-supplemented (medium II plus 1 mM lysine hydroxamate) minimal media; V, thiosine-supplemented (medium II plus 1 mM thiosine) minimal medium.

Figure 23 shows that a highly conserved domain is present in the 5'-UTR of certain gram-positive and gram-negative bacterial mRNAs. Depicted is an alignment of 32 representative mRNA domains from bacteria that conform to the G box consensus sequence. Regions shaded orange, blue and purple identify base-pairing potential of stems P1, P2, and P3, respectively. Nucleotides in red are conserved in greater than 90% of the examples. The asterisk identifies the representative (*xpt-pbuX* 5'-UTR) that was examined in this study. It is important to note that three representatives (BS5, CP4 and VV1) that carry a C to U mutation in the conserved core (in the P3-P1 junction) appear to be adenine-specific riboswitches (unpublished observations). Gene names are as annotated in GenBank, the SubtiList database, or based on protein similarity searches (brackets). Organisms abbreviations are as follows: *Bacillus halodurans* (BH), *Bacillus subtilis* (BS), *Clostridium acetobutylicum* (CA), *Clostridium perfringens* (CP), *Fusobacterium nucleatum* (FN), *Lactococcus lactis* (LL), *Listeria monocytogenes* (LM), *Oceanobacillus iheyensis* (OI), *Staphylococcus aureus* (SA), *Staphylococcus epidermidis* (SE), *Streptococcus agalactiae* (STA), *Streptococcus pyogenes* (STPY), *Streptococcus pneumoniae* (STPN), *Thermoanaerobacter tengcongensis* (TE), and *Vibrio vulnificus* (VV).

Figures 24A, 24B and 24C show the G box RNA of the *xpt-pbuX* mRNA in *B. subtilis* responds allosterically to guanine. Figure 24A shows the consensus sequence and secondary model for the G box RNA domain that resides in the 5' UTR of genes that are largely involved in purine metabolism. Phylogenetic analysis is consistent with the

formation of a three-stem (P1 through P3) junction. Nucleotides depicted in red and black are present in greater than 90% and 80% of the representatives examined (Figure 23). Encircled nucleotides exhibit base complementation, which might indicate the formation of a pseudoknot. Figure 24B shows sequence and ligand-induced structural alterations of the 5'-UTR of the *xpt-pbuX* transcriptional unit. The putative anti-terminator interaction is highlighted in orange. Nucleotides that undergo structural alteration as determined by in-line probing (from C) are identified with red circles. The 93 *xpt* fragment (boxed) of the 201 *xpt* RNA retains guanine-binding function. Asterisks denote alterations to the RNA sequence that facilitate *in vitro* transcription (5' terminus) or that generate a restriction site (3' terminus). Nucleotide numbers begin at the first nucleotide of the natural transcription start site. The translation start codon begins at position 186. Figure 24C shows guanine and related purines selectively induce structural modulation of the 93 *xpt* mRNA fragment. Precursor RNAs (Pre; 5' ³²P-labeled) were subjected to in-line probing by incubation for 40 hr in the absence (-) or presence of guanine, hypoxanthine, xanthine and adenine as indicated by G, H, X and A, respectively. Lanes designated NR, T1 and OH contain RNA that was not reacted, subjected to partial digestion with RNase T1 (G-specific cleavage), or subjected to partial alkaline digestion, respectively. Selected bands corresponding to G-specific cleavage are identified. Regions 1 through 4 identify major sites of ligand-induced modulation of spontaneous RNA cleavage.

Figures 25A and 25B show the 201 *xpt* mRNA Leader Binds Guanine with High Affinity. Figure 25A shows in-line probing reveals that spontaneous RNA cleavage of the 201 *xpt* RNA at four regions decreases with increasing guanine concentrations. Only those locations of the PAGE image corresponding to the four regions of modulation as indicated in Figure 25C are depicted. Other details and notations are as described in the legend to Figure 25C. Figure 25B shows a plot depicting the normalized fraction of RNA that experienced spontaneous cleavage versus the concentration of guanine for modulated regions 1 through 4 in Figure 25A. Fraction cleaved values were normalized to the maximum cleavage measured in the absence of guanine and to the minimum cleavage measured in the presence of 10 μ M guanine. The apparent K_D value (less than or equal to 5 nM) reflects the limits of detection for these assay conditions.

Figures 26A, 26B and 26C show a molecular discrimination by the guanine-binding aptamer of the *xpt-pbuX* mRNA. Figure 26A shows the chemical structures and

apparent K_D values for guanine, hypoxanthine and xanthine (active natural regulators of *xpt-pbuX* genetic expression in *B. subtilis*) versus that of adenine (inactive). Differences in chemical structure relative to guanine are shaded pink. K_D values were established as shown in Figure 26 with the 201 *xpt* RNA. Numbers on guanine represent the positions of the ring nitrogen atoms. Figure 26B shows chemical structures and K_D values for various analogs of guanine reveal that all alterations of this purine cause a loss of binding affinity. Open circles identify K_D values that most likely are significantly higher than indicated, as concentrations of analog above 500 μ M were not examined in this analysis. The apparent K_D values of G, H, X and A as indicated are plotted as red triangles for comparison. Figure 26C shows a schematic representation of the molecular recognition features of the guanine aptamer in 201 *xpt*. Hydrogen bond formation at position 9 of guanine is expected because guanosine ($K_D > 100 \mu$ M) and inosine ($K_D > 100 \mu$ M), which are 9-ribosyl derivatives of guanine and hypoxanthine, respectively, do not exhibit measurable binding (see Figure 27).

Figures 27A and 27B show confirmation of guanine binding specificity by equilibrium dialysis. Figure 27A shows an equilibrium dialysis strategy was used to confirm that *in vitro*-transcribed 93 *xpt* RNAs bind to guanine and can discriminate against various analogs. Each data point was generated by adding 3 H-guanine to chamber *a*, which is separated from RNA and other analogs by a dialysis membrane with a molecular weight cut-off (MWCO) of 5,000 daltons. Left: If no guanine binding sites are present in chamber *b*, or if an excess of unlabeled competitor is present, then no shift in the distribution of tritium is expected. Right: If an excess of guanine-binding RNAs are present in chamber *b*, and if no competitor is present, then a substantial shift in the distribution of tritium towards chamber *b* is expected. Figure 27B shows the 93 *xpt* RNA can shift the distribution of 3 H-guanine in an equilibrium dialysis apparatus, while analogs of guanine are poor competitors. The plot depicts the fraction of counts per minute (cpm) of tritium in chamber *b* relative to the total amount of cpm counted from both chambers. A value of ~0.5 is expected if no shift occurs, as is the case when RNA is absent (none), or in the presence of excess unlabeled competitor (G). A value approaching 1 is expected if the majority of 3 H-guanine is bound by the RNA in chamber *b* in the absence (–) of unlabeled analog, or in the presence of unlabeled analogs that do not serve as effective competitors under the assay conditions (100 nM 3 H-guanine, 300 nM RNA, 500 nM analog). Ino and Gua represents inosine and guanosine, respectively.

Figures 28A, 28B, 28C and 28D show the binding and genetic control functions of variant guanine riboswitches. Figure 28A shows mutations used to examine the importance of various structural features of the guanine aptamer domain. Figure 28B shows examination of the binding function of aptamer variants by equilibrium dialysis. WT designates the wild-type 93 *xpt* construct. Details are as described for Figure 27. Figure 28C shows genetic modulation of a β -galactosidase reporter gene upon the introduction of various purines as indicated. Figure 28D shows regulation of β -galactosidase reporter gene expression by WT and mutants M1 through M7. Open and filled bars represent enzyme activity generated when growing cells in the absence and presence of guanine, respectively.

Figures 29A and 29B show that riboswitches participate in fundamental genetic control. Figure 29A schematic representations of the seven known riboswitches and the metabolites they sense. The secondary structure models were obtained as follows: coenzyme B₁₂ (see Example 1); TPP (see Example 2); FMN (see Example 3), SAM (see Example 7); guanine (see Example 6); lysine (see Example 5); adenine (see Example 8). Coenzyme B₁₂ is depicted in exploded form wherein *a*, *b* and *c* designate covalent attachment sites between fragments. Figure 29B shows a genetic map of *B. subtilis* riboswitch regulons and their positions on the bacterial chromosome. Genes are controlled by riboswitches as identified by matching color. All nomenclature is derived from the SubtiList database release R16.1 (Moszer, I., et al., 1995, Microbiol. 141, 261-268) except for *metI* and *metC*, which are recent designations (Auger, S., et al., 2002, Microbiol. 148, 507-518).

Figures 30A, 30B and 30C show the S Box is a structured RNA domain that binds SAM. (A) Consensus sequence and secondary-structure model of the S box domain derived from 107 bacterial representatives. Red and black positions identify nucleotides whose identity as depicted is conserved in greater than 90% or 80% of the representative S box RNAs, respectively. R, Y, and N represent purine, pyrimidine, and any nucleotide, respectively. P1 through P4 identify conserved base pairing. Encircled nucleotides identify a putative pseudoknot interaction. Figure 30B shows a sequence and secondary structure model for the 251 *yitJ* mRNA fragment. Sites of structural modulation upon introduction of SAM are depicted as described. Nucleotide 1 corresponds to the putative transcriptional start site. Asterisks identify nucleotides that

were added to the construct to permit efficient transcription *in vitro*. The first nucleotide of the AUG start codon is 212 (not shown). Other notations are as described in *a*. Figure 30C shows the spontaneous cleavage patterns of 251 *yitJ* (~1 nM 5' ³²P-labeled) RNA incubated for ~40 hr at 25°C in 50 mM Tris-HCl (pH 8.3 at 25°C), 20 mM MgCl₂, 100 mM KCl, and without (–) or with methionine or SAM as indicated for each lane. NR, T1 and [–]OH represent no reaction, partial digest with RNase T1, and partial digest with alkali, respectively. Certain fragment bands corresponding to T1 digestion (cleaves after G residues) are depicted. Blue arrowheads identify positions of significant modulation of spontaneous cleavage, and the numbered sites were used for quantitation (see Fig. 31*b*). Experimental procedures are similar to those described Examples 1-3.

Figures 31A, 31B and 31C show the binding affinity and molecular discrimination by a SAM-binding RNA. Figure 31A shows the chemical structures of various compounds used to probe the binding characteristics of the SAM *yitJ* riboswitch. Other than methionine, each compound as depicted is coupled to an adenosyl moiety ([A]; inset) coupled via the 5' carbon (as signified by R). Figure 31B Left: The *K_D* of 251 *yitJ* for SAM was determined by plotting the normalized fraction of RNA cleaved at regions 1 through 6 (see Fig. 30*c*) versus the logarithm of the concentration of SAM in molar units. The dashed line indicates the concentration needed to induce half maximal modulation of cleavage activity. Right: *K_D* values for SAM and various analogs as determined by this method. Figure 31C shows molecular discrimination determined by equilibrium dialysis. Assays employed 100 nM of *S*-adenosyl-L-methionine-methyl-³H (³H-SAM; 14.5 μCi mmol^{–1}; ~7,000 cpm) added to side *A* of an equilibrium dialysis chamber (1, 2), and were conducted in the absence (none) or the presence of 3 μM RNA on the *B* side of the chamber as indicated. Equilibrations were carried out for ~10 hr in the absence (–) of unlabeled analogs, and then were subsequently incubated in the presence of 25 μM unlabeled compounds (added to side *B*) as indicated. M1 is a variant of 124 *yitJ* that carries disruptive mutations in the junction between stems P1 and P2 (Fig. 32*a*). Line at a cpm ratio of 1 identifies the bar height expected if a shift in ³H-SAM has not occurred. Additional experimental details are similar to those described in Examples 1 and 2.

Figures 32A, 32B and 32C show the effects of RNA mutations on SAM binding and genetic control. Figure 32A shows the sequence and secondary structure model for the 124 *yitJ* RNA. Mutations M1 through M9 were generated in plasmids containing

fusions of the *yitJ* 5'-UTR upstream from a *lacZ* reporter gene. Templates for preparation of mutant RNAs for *in vitro* studies were then created by PCR, and the mutant DNA constructs were integrated into the chromosome for *in vivo* studies. See Methods for experimental details. Figure 32B shows the analysis of SAM-binding function by equilibrium dialysis in the presence of wild-type (WT) and mutant RNAs as denoted. Details are described in the legend to Fig. 31c, except that 300 nM RNA was used and all assays were conducted without the addition of unlabeled analogs. Figure 32C shows *In vivo* control of β -galactosidase expression in *B. subtilis* cells transformed with various riboswitch constructs as indicated. β -galactosidase activities were measured as described in Example 2. Cells were grown in glucose minimal media in 0.75 $\mu\text{g mL}^{-1}$ methionine (–) 50 $\mu\text{g mL}^{-1}$ methionine (+). M6 through M9 were not examined *in vivo*.

Figures 33A and 33B show metabolite-induced transcription termination of several mRNAs that carry a SAM riboswitch. Figure 33A shows *In vitro* transcription using T7 RNA polymerase results in increased termination of four mRNA leader sequences. Reactions were conducted in the absence (–) or presence (+) of 50 μM of the effector as indicated for each lane. For example, the *metI* template includes the 5' UTR and coding sequences through mRNA position 242, while the termination site is expected to occur at position 189. Below each gel is indicated the percentage of transcription termination (T) at the expected location relative the total amount of expected termination plus full length RNA (FL). Figure 33B shows sequence and structural model for the *metI* riboswitch in two structural states. Green and pink residues correspond to the P1 (anti-anti-terminator) and the terminator stems, respectively. The orange residues correspond to the anti-terminator stem. Sequences boxed in red define the location and identity of mutations used to examine the proposed mechanism of genetic control. Gel: Analysis of mutant *metI* riboswitches wherein disruptive (*Ma*, *Mab* and *Mc*) or the corresponding compensatory mutations (*Mabc*) have been inserted. The *metI* mutant templates and wild-type control template (WT) are identical to the templates used in A, except that the FL product is 220 nucleotides. Other notations are as describe in A.

Figures 34A and 34B show Bacilli species *subtilis* and *anthracis* bind SAM with different affinities. Figure 34A shows structural modulation of the *B. subtilis* *cysH* aptamer as determined by in-line probing. Inset: Apparent K_D values determined by monitoring structural modulation over a range of SAM or SAM analog concentrations.

Two G residues (asterisks) were included at the 5' terminus of the RNA construct to facilitate *in vitro* transcription. Nucleotide numbers are given relative to the putative transcription start site. In-line probing was conducted with an RNA extending to nucleotide 117, while the remainder of the RNA is shown to depict the putative transcription terminator stem. Experiments were similar to those described in Fig. 30b and Fig. 31b. See the legend for Fig. 30b for details. Figure 34B shows structural modulation of the *B. subtilis* *cysH* aptamer as determined by in-line probing. The transcription start point of the *B. anthracis* *cysH* mRNA has not been determined, and so numbering of nucleotides begins immediately after the two inserted G residues (asterisks). In-line probing was conducted with an RNA extending to nucleotide 112. See *a* for additional details.

Figures 35A, 35B and 35C show guanine- and adenine-specific riboswitches. Figure 35a shows sequence and structural features of the two guanine-specific (*purE* and *xpt*) and three adenine-specific aptamer domains that are examined in this study. P1 through P3 identify the three base-paired stems comprising the secondary structure of the aptamer domain. Red nucleotides identify positions whose base identity is conserved in greater than 90% of representatives in the phylogeny¹. The arrow identifies a nucleotide within the conserved core of the aptamer that is a determinant of ligand specificity. BS, CP and VV designate *B. subtilis*, *Clostridium perfringens* and *Vibrio vulnificus*, respectively. Figure 35b shows sequence and secondary structure of the *xpt* and *ydhL* aptamers. Green nucleotides identify positions within the *ydhL* aptamer that differ from those in the *xpt* aptamer. Nucleotides in *xpt* are numbered as described in Example 6. Other notations are as described in A.

Figures 36A, 36B, 36C, 36D and 36E show the ligand specificity of five G box RNAs. (a through e) In-line probing assays for the conserved aptamer domains as labeled. NR, T1 and ⁻OH identify marker lanes wherein precursor RNAs (Pre) were not incubated, or were partially digested with RNase T1 or alkali, respectively. Selected bands corresponding to RNase T1 digestion (cleavage 3' relative to guanidyl residues) are labeled for each RNA. RNAs were incubated for 40 hr in the absence of ligand (–), or in the presence of 1 μ M guanine (G) or adenine (A). Large arrowheads identify sites of substantial change in cleavage pattern that is due to the addition of a particular ligand. See Methods for additional details.

Figures 37A and 37 B show the binding affinity of the *ydhl* aptamer for adenine. Figure 37a shows the in-line probing assay for the 80 *ydhl* RNA at various concentrations of adenine. For each lane, sites 1 through 4 were quantitated and the fraction of RNA cleaved was used to determine the apparent K_D . Figure 37b shows a plot of the normalized fraction of RNA that has undergone spontaneous cleavage at sites 1 through 4 versus the concentration of adenine. See Example 8 for additional details.

Figures 38A and 38B show the specificity of molecular recognition by the adenine aptamer from *ydhl*. Figure 38a Top: Chemical structures of adenine, guanine and other purine analogs that exhibit measurable binding to the 80 *ydhl* RNA. Chemical changes relative to 2,6-DAP, which is the tightest-binding compound, are highlighted in pink. Bottom left: Plot of the apparent K_D values for various purines. Bottom right: Model for the chemical features on adenine that serve as molecular recognition contacts for *ydhl*. Note that the importance of N7 and N9 has not been determined. Encircled arrow indicated that a contact could exist if a hydrogen bond donor is appended to C2. Figure 38b shows chemical structures of various purines that are not bound by the 80 *ydhl* RNA (K_D values poorer than 300 μ M).

Figures 39A, 39B, 39C and 39D show interconversion of guanine- and adenine-specific aptamers. Figure 39a Left: Plot of the normalized fraction of wild-type 93 *xpt* RNA cleavage product for a given site versus the logarithm of the concentration of ligand present during incubation in an in-line probing assay. Cleavage products monitored for modulation correspond to site 3 (Fig. 37a). Right: Plot of the fraction of the total counts per minute (cpm) present in chamber B relative to the total counts per minute from sides A and B of an equilibrium dialysis chamber. Value of ~0.5 indicate an equal distribution of ligand (no binding) while values of ~1 indicate that most of the ligand is bound to the RNA within side B of the chamber. (b, c, d) In-line probing plots and equilibrium dialysis plots for 93 *xpt* (C to U mutation), 80 *ydhl*, and 80 *ydhl* (U to C mutation), respectively. Details are describe in a, or are described in the Example 8.

Figures 40A, 40B, 40C, 40D and 40E show a model for the genetic control of *ydhl* by an adenine riboswitch and its function as a gene-activating element. Figure 40a sequence of the adenine riboswitch from *B. subtilis ydhl* and secondary structure models for the 'ON' and 'OFF' states for gene regulation. Figure 40b *In vivo* function of the

wild-type *ydhL* riboswitch and of a variant form as determined by fusion to a β -galactosidase reporter gene.

DETAILED DESCRIPTION OF THE INVENTION

The disclosed methods and compositions can be understood more readily by
5 reference to the following detailed description of particular embodiments and the Example included therein and to the Figures and their previous and following description.

Certain natural mRNAs serve as metabolite-sensitive genetic switches wherein the RNA directly binds a small organic molecule. This binding process changes the
10 conformation of the mRNA, which causes a change in gene expression by a variety of different mechanisms. Modified versions of these natural "riboswitches" (created by using various nucleic acid engineering strategies) can be employed as designer genetic switches that are controlled by specific effector compounds (referred to herein as trigger molecules). The natural switches are targets for antibiotics and other small molecule
15 therapies. In addition, the architecture of riboswitches allows actual pieces of the natural switches to be used to construct new non-immunogenic genetic control elements, for example the aptamer (molecular recognition) domain can be swapped with other non-natural aptamers (or otherwise modified) such that the new recognition domain causes genetic modulation with user-defined effector compounds. The changed switches
20 become part of a therapy regimen - turning on, or off, or regulating protein synthesis. Newly constructed genetic regulation networks can be applied in such areas as living biosensors, metabolic engineering of organisms, and in advanced forms of gene therapy treatments.

Messenger RNAs are typically thought of as passive carriers of genetic
25 information that are acted upon by protein- or small RNA-regulatory factors and by ribosomes during the process of translation. It was discovered that certain mRNAs carry natural aptamer domains and that binding of specific metabolites directly to these RNA domains leads to modulation of gene expression. Natural riboswitches exhibit two surprising functions that are not typically associated with natural RNAs. First, the
30 mRNA element can adopt distinct structural states wherein one structure serves as a precise binding pocket for its target metabolite. Second, the metabolite-induced allosteric interconversion between structural states causes a change in the level of gene expression

by one of several distinct mechanisms. Riboswitches typically can be dissected into two separate domains: one that selectively binds the target (aptamer domain) and another that influences genetic control (expression platform). It is the dynamic interplay between these two domains that results in metabolite-dependent allosteric control of gene
5 expression.

As disclosed herein, distinct classes of riboswitches have been identified and are shown to selectively recognize activating compounds (referred to herein as trigger molecules). For example, coenzyme B₁₂, thiamine pyrophosphate (TPP), and flavin mononucleotide (FMN) activate riboswitches present in genes encoding key enzymes in
10 metabolic or transport pathways of these compounds. The aptamer domain of each riboswitch class conforms to a highly conserved consensus sequence and structure. Thus, sequence homology searches can be used to identify related riboswitch domains. Riboswitch domains have been discovered in various organisms from bacteria, archaea, and eukarya.

One class of riboswitches that recognizes guanine and discriminates against most other purine analogs has been discovered. Representative RNAs that carry the consensus sequence and structural features of guanine riboswitches are located in the 5'-
15 untranslated region (UTR) of numerous genes of prokaryotes, where they control expression of proteins involved in purine salvage and biosynthesis. Three representatives of this phylogenetic collection bind adenine with values for apparent dissociation
20 constant (apparent K_D) that are several orders of magnitude better than for guanine. The preference for adenine is due to a single nucleotide substitution in the core of the riboswitch, wherein each representative most likely recognizes its corresponding ligand by forming a Watson/Crick base pair. In addition, the adenine-specific riboswitch
25 associated with the *ydhL* gene of *Bacillus subtilis* functions as a genetic 'ON' switch, wherein adenine binding causes a structural rearrangement that precludes formation of an intrinsic transcription terminator stem. Guanine-sensing riboswitches are a class of RNA genetic control elements that modulate gene expression in response to changing
concentrations of this compound.

It was discovered that the 5'-untranslated sequence of the *Escherichia coli* *btuB*
30 mRNA assumes a more proactive role in metabolic monitoring and genetic control. The mRNA serves as a metabolite-sensing genetic switch by selectively binding coenzyme B₁₂ without the need for proteins. This binding event establishes a distinct RNA

structure that is likely to be responsible for inhibition of ribosome binding and consequent reduction in synthesis of the cobalamin transport protein BtuB. This discovery, along with related observations described herein, supports the hypothesis that metabolic monitoring through RNA-metabolite interactions is a widespread mechanism of genetic control.

RNA structure probing data indicate that the thiamine pyrophosphate (TPP) riboswitch operates as an allosteric sensor of its target compound, wherein binding of TPP by the aptamer domain stabilizes a conformational state within the aptamer and within the neighboring expression platform that precludes translation. The diversity of expression platforms appears to be expansive. The *thiM* RNA uses a Shine-Dalgarno (SD)-blocking mechanism to control translation. In contrast, the *thiC* RNA controls gene expression both at transcription and translation, and therefore might make use of a somewhat more complex expression platform that converts the TPP binding event into a transcription termination event and into inhibition of translation of completed mRNAs.

1. General Organization of Riboswitch RNAs

Bacterial riboswitch RNAs are genetic control elements that are located primarily within the 5'-untranslated region (5'-UTR) of the main coding region of a particular mRNA. Structural probing studies (discussed further below) reveal that riboswitch elements are generally composed of two domains: a natural aptamer (T. Hermann, D. J. Patel, *Science* 2000, 287, 820; L. Gold, et al., *Annual Review of Biochemistry* 1995, 64, 763) that serves as the ligand-binding domain, and an 'expression platform' that interfaces with RNA elements that are involved in gene expression (e.g. Shine-Dalgarno (SD) elements; transcription terminator stems). These conclusions are drawn from the observation that aptamer domains synthesized *in vitro* bind the appropriate ligand in the absence of the expression platform (see Examples 2, 3 and 6). Moreover, structural probing investigations suggest that the aptamer domain of most riboswitches adopts a particular secondary- and tertiary-structure fold when examined independently, that is essentially identical to the aptamer structure when examined in the context of the entire 5' leader RNA. This implies that, in many cases, the aptamer domain is a modular unit that folds independently of the expression platform (see Examples 2, 3 and 6).

Ultimately, the ligand-bound or unbound status of the aptamer domain is interpreted through the expression platform, which is responsible for exerting an influence upon gene expression. The view of a riboswitch as a modular element is further

supported by the fact that aptamer domains are highly conserved amongst various organisms (and even between kingdoms as is observed for the TPP riboswitch), (N. Sudarsan, et al., *RNA* 2003, 9, 644) whereas the expression platform varies in sequence, structure, and in the mechanism by which expression of the appended open reading frame is controlled. For example, ligand binding to the TPP riboswitch of the *tenA* mRNA of *B. subtilis* causes transcription termination (A. S. Mironov, et al., *Cell* 2002, 111, 747). This expression platform is distinct in sequence and structure compared to the expression platform of the TPP riboswitch in the *thiM* mRNA from *E. coli*, wherein TPP binding causes inhibition of translation by a SD blocking mechanism (see Example 2).

10 The TPP aptamer domain is easily recognizable and of near identical functional character between these two transcriptional units, but the genetic control mechanisms and the expression platforms that carry them out are very different.

Aptamer domains for riboswitch RNAs typically range from ~70 to 170 nt in length (Figure 11). This observation was somewhat unexpected given that *in vitro* evolution experiments identified a wide variety of small molecule-binding aptamers, which are considerably shorter in length and structural intricacy (T. Hermann, D. J. Patel, *Science* 2000, 287, 820; L. Gold, et al., *Annual Review of Biochemistry* 1995, 64, 763; M. Famulok, *Current Opinion in Structural Biology* 1999, 9, 324). Although the reasons for the substantial increase in complexity and information content of the natural aptamer sequences relative to artificial aptamers remains to be proven, this complexity is most likely required to form RNA receptors that function with high affinity and selectivity. Apparent K_D values for the ligand-riboswitch complexes range from low nanomolar to low micromolar. It is also worth noting that some aptamer domains, when isolated from the appended expression platform, exhibit improved affinity for the target ligand over that of the intact riboswitch. (~10 to 100-fold) (see Example 2). Presumably, there is an energetic cost in sampling the multiple distinct RNA conformations required by a fully intact riboswitch RNA, which is reflected by a loss in ligand affinity. Since the aptamer domain must serve as a molecular switch, this might also add to the functional demands on natural aptamers that might help rationalize their more sophisticated structures.

2. Riboswitch Regulation of Transcription Termination in Bacteria

Bacteria primarily make use of two methods for termination of transcription. Certain genes incorporate a termination signal that is dependent upon the Rho protein, (J.

P. Richardson, *Biochimica et Biophysica Acta* 2002, 1577, 251). while others make use of Rho-independent terminators (intrinsic terminators) to destabilize the transcription elongation complex (I. Gusarov, E. Nudler, *Molecular Cell* 1999, 3, 495; E. Nudler, M. E. Gottesman, *Genes to Cells* 2002, 7, 755). The latter RNA elements are composed of a GC-rich stem-loop followed by a stretch of 6-9 uridyl residues. Intrinsic terminators are widespread throughout bacterial genomes (F. Lillo, et al., 2002, 18, 971), and are typically located at the 3'-termini of genes or operons. Interestingly, an increasing number of examples are being observed for intrinsic terminators located within 5'-UTRs.

Amongst the wide variety of genetic regulatory strategies employed by bacteria there is a growing class of examples wherein RNA polymerase responds to a termination signal within the 5'-UTR in a regulated fashion (T. M. Henkin, *Current Opinion in Microbiology* 2000, 3, 149). During certain conditions the RNA polymerase complex is directed by external signals either to perceive or to ignore the termination signal. Although transcription initiation might occur without regulation, control over mRNA synthesis (and of gene expression) is ultimately dictated by regulation of the intrinsic terminator. Presumably, one of at least two mutually exclusive mRNA conformations results in the formation or disruption of the RNA structure that signals transcription termination. A trans-acting factor, which in some instances is a RNA (F. J. Grundy, et al., *Proceedings of the National Academy of Sciences of the United States of America* 2002, 99, 11121; T. M. Henkin, C. Yanofsky, *Bioessays* 2002, 24, 700) and in others is a protein (J. Stulke, *Archives of Microbiology* 2002, 177, 433), is generally required for receiving a particular intracellular signal and subsequently stabilizing one of the RNA conformations. Riboswitches offer a direct link between RNA structure modulation and the metabolite signals that are interpreted by the genetic control machinery. A brief overview of the FMN riboswitch from a *B. subtilis* mRNA is provided below to illustrate this mechanism.

It was discovered that certain mRNAs involved in thiamine biosynthesis bind to thiamine (vitamin B₁) or its bioactive pyrophosphate derivative (TPP) without the participation of protein factors. The mRNA-effector complex adopts a distinct structure that sequesters the ribosome-binding site and leads to a reduction in gene expression. This metabolite-sensing mRNA system provides an example of a genetic "riboswitch" (referred to herein as a riboswitch) whose origin might predate the evolutionary emergence of proteins. It has been discovered that the mRNA leader sequence of the

btuB gene of *Escherichia coli* can bind coenzyme B₁₂ selectively, and that this binding event brings about a structural change in the RNA that is important for genetic control (see Example 1). It was also discovered that mRNAs that encode thiamine biosynthetic proteins also employ a riboswitch mechanism (see Example 2).

5 It was also discovered that the 5'-UTR of the *lysC* gene of *Bacillus subtilis* carries a conserved RNA element that serves as a lysine-responsive riboswitch. The ligand-binding domain of the riboswitch binds to L-lysine with an apparent dissociation constant (K_D) of approximately 1 μ M, and exhibits a high level of molecular discrimination against closely related analogs including D-lysine and ornithine. This
10 widespread class of riboswitches serves as a target for the antimicrobial agent thiosine.

 It was also discovered that the *xpt-pbuX* operon (Christiansen, L.C., et al., 1997, *J. Bacteriol.* 179, 2540-2550) is controlled by a riboswitch that exhibits high affinity and high selectivity for guanine. This class of riboswitches is present in the 5'-untranslated region (5'-UTR) of five transcriptional units in *B. subtilis*, including that of the 12-gene
15 *pur* operon. Direct binding of guanine by mRNAs serves as a critical determinant of metabolic homeostasis for purine metabolism in certain bacteria. Furthermore, the discovered classes of riboswitches, which respond to seven distinct target molecules, control at least 68 genes in *Bacillus subtilis* that are of fundamental importance to central metabolic pathways.

20 It was discovered that a highly conserved RNA domain termed the S box serves as a selective aptamer for SAM. Allosteric modulation of secondary and tertiary structures are induced upon SAM binding to the aptamer domain, and these structural changes are responsible for inducing termination of mRNA transcription.

 A variant class of riboswitches that responds to adenine is also disclosed. These
25 riboswitches carry an aptamer domain that corresponds closely in sequence and secondary structure to the guanine aptamer. However, each representative of the adenine sub-class of riboswitches carries a C to U mutation in the conserved core of the aptamer, indicating that this residue is involved in metabolite recognition. The identity of this single nucleotide determines the binding specificity between guanine and adenine, which
30 provides an example of how complex riboswitch structures can be mutated to recognize new metabolite targets.

Although the specific natural riboswitches disclosed herein are the first examples of mRNA elements that control genetic expression by metabolite binding, it is expected that this genetic control strategy is widespread in biology. It has been suggested (White III, Coenzymes as fossils of an earlier metabolic state. *J. Mol. Evol.* 7, 101-104 (1976);
5 White III, In: *The Pyridine Nucleotide Coenzymes*. Acad. Press, NY pp. 1-17 (1982); Benner et al., Modern metabolism as a palimpsest of the RNA world. *Proc. Natl. Acad. Sci. USA* 86, 7054-7058 (1989)) that TPP, coenzyme B₁₂ and FMN emerged as biological cofactors during the RNA world (Joyce, The antiquity of RNA-based evolution. *Nature* 418, 214-221 (2002)). If these metabolites were being biosynthesized
10 and used before the advent of proteins, then certain riboswitches might be modern examples of the most ancient form of genetic control. A search of genomic sequence databases has revealed that sequences corresponding to the TPP aptamer exist in organisms from bacteria, archaea and eukarya—largely without major alteration. Although new metabolite-binding mRNAs are likely to emerge as evolution progresses,
15 it is possible that the known riboswitches are molecular fossils from the RNA world.

Disclosed are mRNA elements that have been identified in fungi and in plants that match the consensus sequence and structure of thiamine pyrophosphate-binding domains of prokaryotes. In *Arabidopsis*, the consensus motif resides in the 3'-UTR of a thiamine biosynthetic gene, and the isolated RNA domain binds the corresponding
20 coenzyme in vitro. These results indicate that metabolite-binding mRNAs are involved in eukaryotic gene regulation and that some riboswitches might be representatives of an ancient form of genetic control.

It is to be understood that the disclosed method and compositions are not limited to specific synthetic methods, specific analytical techniques, or to particular reagents
25 unless otherwise specified, and, as such, can vary. It is also to be understood that the terminology used herein is for the purpose of describing particular embodiments only and is not intended to be limiting.

Materials

Disclosed are materials, compositions, and components that can be used for, can
30 be used in conjunction with, can be used in preparation for, or are products of the disclosed methods and compositions. These and other materials are disclosed herein, and it is understood that when combinations, subsets, interactions, groups, etc. of these materials are disclosed that while specific reference to each of various individual and

collective combinations and permutation of these compounds can not be explicitly disclosed, each is specifically contemplated and described herein. For example, if a riboswitch or aptamer domain is disclosed and discussed and a number of modifications that can be made to a number of molecules including the riboswitch or aptamer domain are discussed, each and every combination and permutation of riboswitch or aptamer domain and the modifications that are possible are specifically contemplated unless specifically indicated to the contrary. Thus, if a class of molecules A, B, and C are disclosed as well as a class of molecules D, E, and F and an example of a combination molecule, A-D is disclosed, then even if each is not individually recited, each is individually and collectively contemplated. Thus, in this example, each of the combinations A-E, A-F, B-D, B-E, B-F, C-D, C-E, and C-F are specifically contemplated and should be considered disclosed from disclosure of A, B, and C; D, E, and F; and the example combination A-D. Likewise, any subset or combination of these is also specifically contemplated and disclosed. Thus, for example, the sub-group of A-E, B-F, and C-E are specifically contemplated and should be considered disclosed from disclosure of A, B, and C; D, E, and F; and the example combination A-D. This concept applies to all aspects of this application including, but not limited to, steps in methods of making and using the disclosed compositions. Thus, if there are a variety of additional steps that can be performed it is understood that each of these additional steps can be performed with any specific embodiment or combination of embodiments of the disclosed methods, and that each such combination is specifically contemplated and should be considered disclosed.

A. Riboswitches

Riboswitches are expression control elements that are part of the RNA molecule to be expressed and that change state when bound by a trigger molecule. Riboswitches typically can be dissected into two separate domains: one that selectively binds the target (aptamer domain) and another that influences genetic control (expression platform domain). It is the dynamic interplay between these two domains that results in metabolite-dependent allosteric control of gene expression. Disclosed are isolated and recombinant riboswitches, recombinant constructs containing such riboswitches, heterologous sequences operably linked to such riboswitches, and cells and transgenic organisms harboring such riboswitches, riboswitch recombinant constructs, and riboswitches operably linked to heterologous sequences. The heterologous sequences

can be, for example, sequences encoding proteins or peptides of interest, including reporter proteins or peptides. Preferred riboswitches are, or are derived from, naturally occurring riboswitches.

5 The disclosed riboswitches, including the derivatives and recombinant forms thereof, generally can be from any source, including naturally occurring riboswitches and riboswitches designed de novo. Any such riboswitches can be used in or with the disclosed methods. However, different types of riboswitches can be defined and some such sub-types can be useful in or with particular methods (generally as described elsewhere herein). Types of riboswitches include, for example, naturally occurring
10 riboswitches, derivatives and modified forms of naturally occurring riboswitches, chimeric riboswitches, and recombinant riboswitches. A naturally occurring riboswitch is a riboswitch having the sequence of a riboswitch as found in nature. Such a naturally occurring riboswitch can be an isolated or recombinant form of the naturally occurring riboswitch as it occurs in nature. That is, the riboswitch has the same primary structure
15 but has been isolated or engineered in a new genetic or nucleic acid context. Chimeric riboswitches can be made up of, for example, part of a riboswitch of any or of a particular class or type of riboswitch and part of a different riboswitch of the same or of any different class or type of riboswitch; part of a riboswitch of any or of a particular class or type of riboswitch and any non-riboswitch sequence or component.
20 Recombinant riboswitches are riboswitches that have been isolated or engineered in a new genetic or nucleic acid context.

Different classes of riboswitches refer to riboswitches that have the same or similar trigger molecules or riboswitches that have the same or similar overall structure (predicted, determined, or a combination). Riboswitches of the same class generally, but
25 need not, have both the same or similar trigger molecules and the same or similar overall structure.

Also disclosed are chimeric riboswitches containing heterologous aptamer domains and expression platform domains. That is, chimeric riboswitches are made up an aptamer domain from one source and an expression platform domain from another
30 source. The heterologous sources can be from, for example, different specific riboswitches, different types of riboswitches, or different classes of riboswitches. The heterologous aptamers can also come from non-riboswitch aptamers. The heterologous expression platform domains can also come from non-riboswitch sources.

Riboswitches can be modified from other known, developed or naturally-occurring riboswitches. For example, switch domain portions can be modified by changing one or more nucleotides while preserving the known or predicted secondary, tertiary, or both secondary and tertiary structure of the riboswitch. For example, both
5 nucleotides in a base pair can be changed to nucleotides that can also base pair. Changes that allow retention of base pairing are referred to herein as base pair conservative changes.

Modified or derivative riboswitches can also be produced using in vitro selection and evolution techniques. In general, in vitro evolution techniques as applied to
10 riboswitches involve producing a set of variant riboswitches where part(s) of the riboswitch sequence is varied while other parts of the riboswitch are held constant. Activation, deactivation or blocking (or other functional or structural criteria) of the set of variant riboswitches can then be assessed and those variant riboswitches meeting the criteria of interest are selected for use or further rounds of evolution. Useful base
15 riboswitches for generation of variants are the specific and consensus riboswitches disclosed herein. Consensus riboswitches can be used to inform which part(s) of a riboswitch to vary for in vitro selection and evolution.

Also disclosed are modified riboswitches with altered regulation. The regulation of a riboswitch can be altered by operably linking an aptamer domain to the expression
20 platform domain of the riboswitch (which is a chimeric riboswitch). The aptamer domain can then mediate regulation of the riboswitch through the action of, for example, a trigger molecule for the aptamer domain. Aptamer domains can be operably linked to expression platform domains of riboswitches in any suitable manner, including, for example, by replacing the normal or natural aptamer domain of the riboswitch with the
25 new aptamer domain. Generally, any compound or condition that can activate, deactivate or block the riboswitch from which the aptamer domain is derived can be used to activate, deactivate or block the chimeric riboswitch.

Also disclosed are inactivated riboswitches. Riboswitches can be inactivated by covalently altering the riboswitch (by, for example, crosslinking parts of the riboswitch
30 or coupling a compound to the riboswitch). Inactivation of a riboswitch in this manner can result from, for example, an alteration that prevents the trigger molecule for the riboswitch from binding, that prevents the change in state of the riboswitch upon binding

of the trigger molecule, or that prevents the expression platform domain of the riboswitch from affecting expression upon binding of the trigger molecule.

Also disclosed are biosensor riboswitches. Biosensor riboswitches are engineered riboswitches that produce a detectable signal in the presence of their cognate trigger molecule. Useful biosensor riboswitches can be triggered at or above threshold levels of the trigger molecules. Biosensor riboswitches can be designed for use in vivo or in vitro. For example, biosensor riboswitches operably linked to a reporter RNA that encodes a protein that serves as or is involved in producing a signal can be used in vivo by engineering a cell or organism to harbor a nucleic acid construct encoding the riboswitch/reporter RNA. An example of a biosensor riboswitch for use in vitro is a riboswitch that includes a conformation dependent label, the signal from which changes depending on the activation state of the riboswitch. Such a biosensor riboswitch preferably uses an aptamer domain from or derived from a naturally occurring riboswitch. Biosensor riboswitches can be used in various situations and platforms. For example, biosensor riboswitches can be used with solid supports, such as plates, chips, strips and wells.

Also disclosed are modified or derivative riboswitches that recognize new trigger molecules. New riboswitches and/or new aptamers that recognize new trigger molecules can be selected for, designed or derived from known riboswitches. This can be accomplished by, for example, producing a set of aptamer variants in a riboswitch, assessing the activation of the variant riboswitches in the presence of a compound of interest, selecting variant riboswitches that were activated (or, for example, the riboswitches that were the most highly or the most selectively activated), and repeating these steps until a variant riboswitch of a desired activity, specificity, combination of activity and specificity, or other combination of properties results.

Particularly useful aptamer domains can form a stem structure referred to herein as the P1 stem structure (or simply P1). The P1 stems of a variety of riboswitches are shown in Figure 11 (and in other figures). The hybridizing strands in the P1 stem structure are referred to as the aptamer strand (also referred to as the P1a strand) and the control strand (also referred to as the P1b strand). The control strand can form a stem structure with both the aptamer strand and a sequence in a linked expression platform that is referred to as the regulated strand (also referred to as the P1c strand). Thus, the control strand (P1b) can form alternative stem structures with the aptamer strand (P1a)

and the regulated strand (P1c). Activation and deactivation of a riboswitch results in a shift from one of the stem structures to the other (from P1a/P1b to P1b/P1c or vice versa). The formation of the P1b/P1c stem structure affects expression of the RNA molecule containing the riboswitch. Riboswitches that operate via this control
5 mechanism are referred to herein as alternative stem structure riboswitches (or as alternative stem riboswitches).

In general, any aptamer domain can be adapted for use with any expression platform domain by designing or adapting a regulated strand in the expression platform domain to be complementary to the control strand of the aptamer domain. Alternatively,
10 the sequence of the aptamer and control strands of an aptamer domain can be adapted so that the control strand is complementary to a functionally significant sequence in an expression platform. For example, the control strand can be adapted to be complementary to the Shine-Dalgarno sequence of an RNA such that, upon formation of a stem structure between the control strand and the SD sequence, the SD sequence
15 becomes inaccessible to ribosomes, thus reducing or preventing translation initiation. Note that the aptamer strand would have corresponding changes in sequence to allow formation of a P1 stem in the aptamer domain.

As another example, a transcription terminator can be added to an RNA molecule (most conveniently in an untranslated region of the RNA) where part of the sequence of
20 the transcription terminator is complementary to the control strand of an aptamer domain (the sequence will be the regulated strand). This will allow the control sequence of the aptamer domain to form alternative stem structures with the aptamer strand and the regulated strand, thus either forming or disrupting a transcription terminator stem upon activation or deactivation of the riboswitch. Any other expression element can be
25 brought under the control of a riboswitch by similar design of alternative stem structures.

For transcription terminators controlled by riboswitches, the speed of transcription and spacing of the riboswitch and expression platform elements can be important for proper control. Transcription speed can be adjusted by, for example, by including polymerase pausing elements (e.g., a series of uridine residues) to pause
30 transcription and allow the riboswitch to form and sense trigger molecules. For example, with the FMN riboswitch, if FMN is bound to its aptamer domain, then the antiterminator sequence is sequestered and is unavailable for formation of an antiterminator structure (Fig. 12). However, if FMN is absent, the antiterminator can

form once its nucleotides emerge from the polymerase. RNAP then breaks free of the pause site only to reach another U-stretch and pause again. The transcriptional terminator then forms only if the terminator nucleotides are not tied up by the antiterminator.

5 Disclosed are regulatable gene expression constructs comprising a nucleic acid molecule encoding an RNA comprising a riboswitch operably linked to a coding region, wherein the riboswitch regulates expression of the RNA, wherein the riboswitch and coding region are heterologous. The riboswitch can comprise an aptamer domain and an expression platform domain, wherein the aptamer domain and the expression platform
10 domain are heterologous. The riboswitch can comprise an aptamer domain and an expression platform domain, wherein the aptamer domain comprises a P1 stem, wherein the P1 stem comprises an aptamer strand and a control strand, wherein the expression platform domain comprises a regulated strand, wherein the regulated strand, the control strand, or both have been designed to form a stem structure.

15 Disclosed are riboswitches, wherein the riboswitch is a non-natural derivative of a naturally-occurring riboswitch. The riboswitch can comprise an aptamer domain and an expression platform domain, wherein the aptamer domain and the expression platform domain are heterologous. The riboswitch can be derived from a naturally-occurring
20 guanine-responsive riboswitch, adenine-responsive riboswitch, lysine-responsive riboswitch, thiamine pyrophosphate-responsive riboswitch, adenosylcobalamin-responsive riboswitch, flavin mononucleotide-responsive riboswitch, or a S-adenosylmethionine-responsive riboswitch. The riboswitch can be activated by a trigger molecule, wherein the riboswitch produces a signal when activated by the trigger molecule.

25 Numerous riboswitches and riboswitch constructs are described and referred to herein. It is specifically contemplated that any specific riboswitch or riboswitch construct or group of riboswitches or riboswitch constructs can be excluded from some aspects of the invention disclosed herein. For example, fusion of the *xpt-pbuX*
30 riboswitch with a reporter gene could be excluded from a set of riboswitches fused to reporter genes.

1. Aptamer Domains

Aptamers are nucleic acid segments and structures that can bind selectively to particular compounds and classes of compounds. Riboswitches have aptamer domains

that, upon binding of a trigger molecule result in a change the state or structure of the riboswitch. In functional riboswitches, the state or structure of the expression platform domain linked to the aptamer domain changes when the trigger molecule binds to the aptamer domain. Aptamer domains of riboswitches can be derived from any source, including, for example, natural aptamer domains of riboswitches, artificial aptamers, engineered, selected, evolved or derived aptamers or aptamer domains. Aptamers in riboswitches generally have at least one portion that can interact, such as by forming a stem structure, with a portion of the linked expression platform domain. This stem structure will either form or be disrupted upon binding of the trigger molecule.

Consensus aptamer domains of a variety of natural riboswitches are shown in Figure 11. These aptamer domains (including all of the direct variants embodied therein) can be used in riboswitches. The consensus sequences and structures indicate variations in sequence and structure. Aptamer domains that are within the indicated variations are referred to herein as direct variants. These aptamer domains can be modified to produce modified or variant aptamer domains. Conservative modifications include any change in base paired nucleotides such that the nucleotides in the pair remain complementary. Moderate modifications include changes in the length of stems or of loops (for which a length or length range is indicated) of less than or equal to 20% of the length range indicated. Loop and stem lengths are considered to be "indicated" where the consensus structure shows a stem or loop of a particular length or where a range of lengths is listed or depicted. Moderate modifications include changes in the length of stems or of loops (for which a length or length range is not indicated) of less than or equal to 40% of the length range indicated. Moderate modifications also include and functional variants of unspecified portions of the aptamer domain. Unspecified portions of the aptamer domains are indicated by solid lines in Figure 11.

The P1 stem and its constituent strands can be modified in adapting aptamer domains for use with expression platforms and RNA molecules. Such modifications, which can be extensive, are referred to herein as P1 modifications. P1 modifications include changes to the sequence and/or length of the P1 stem of an aptamer domain.

The aptamer domains shown in Figure 11 (including any direct variants) are particularly useful as initial sequences for producing derived aptamer domains via in vitro selection or in vitro evolution techniques.

Aptamer domains of the disclosed riboswitches can also be used for any other purpose, and in any other context, as aptamers. For example, aptamers can be used to control ribozymes, other molecular switches, and any RNA molecule where a change in structure can affect function of the RNA.

5 **2. Expression Platform Domains**

Expression platform domains are a part of riboswitches that affect expression of the RNA molecule that contains the riboswitch. Expression platform domains generally have at least one portion that can interact, such as by forming a stem structure, with a portion of the linked aptamer domain. This stem structure will either form or be
10 disrupted upon binding of the trigger molecule. The stem structure generally either is, or prevents formation of, an expression regulatory structure. An expression regulatory structure is a structure that allows, prevents, enhances or inhibits expression of an RNA molecule containing the structure. Examples include Shine-Dalgarno sequences, initiation codons, transcription terminators, and stability and processing signals.

15 **B. Trigger Molecules**

Trigger molecules are molecules and compounds that can activate a riboswitch. This includes the natural or normal trigger molecule for the riboswitch and other compounds that can activate the riboswitch. Natural or normal trigger molecules are the trigger molecule for a given riboswitch in nature or, in the case of some non-natural
20 riboswitches, the trigger molecule for which the riboswitch was designed or with which the riboswitch was selected (as in, for example, in vitro selection or in vitro evolution techniques). Non-natural trigger molecules can be referred to as non-natural trigger molecules.

C. Compounds

25 Also disclosed are compounds, and compositions containing such compounds, that can activate, deactivate or block a riboswitch. Riboswitches function to control gene expression through the binding or removal of a trigger molecule. Compounds can be used to activate, deactivate or block a riboswitch. The trigger molecule for a riboswitch (as well as other activating compounds) can be used to activate a riboswitch.
30 Compounds other than the trigger molecule generally can be used to deactivate or block a riboswitch. Riboswitches can also be deactivated by, for example, removing trigger molecules from the presence of the riboswitch. A riboswitch can be blocked by, for

example, binding of an analog of the trigger molecule that does not activate the riboswitch.

Also disclosed are compounds for altering expression of an RNA molecule, or of a gene encoding an RNA molecule, where the RNA molecule includes a riboswitch.

5 This can be accomplished by bringing a compound into contact with the RNA molecule. Riboswitches function to control gene expression through the binding or removal of a trigger molecule. Thus, subjecting an RNA molecule of interest that includes a riboswitch to conditions that activate, deactivate or block the riboswitch can be used to alter expression of the RNA. Expression can be altered as a result of, for example, 10 termination of transcription or blocking of ribosome binding to the RNA. Binding of a trigger molecule can, depending on the nature of the riboswitch, reduce or prevent expression of the RNA molecule or promote or increase expression of the RNA molecule.

Also disclosed are compounds for regulating expression of an RNA molecule, or 15 of a gene encoding an RNA molecule. Also disclosed are compounds for regulating expression of a naturally occurring gene or RNA that contains a riboswitch by activating, deactivating or blocking the riboswitch. If the gene is essential for survival of a cell or organism that harbors it, activating, deactivating or blocking the riboswitch can in death, stasis or debilitation of the cell or organism.

20 Also disclosed are compounds for regulating expression of an isolated, engineered or recombinant gene or RNA that contains a riboswitch by activating, deactivating or blocking the riboswitch. If the gene encodes a desired expression product, activating or deactivating the riboswitch can be used to induce expression of the gene and thus result in production of the expression product. If the gene encodes an 25 inducer or repressor of gene expression or of another cellular process, activation, deactivation or blocking of the riboswitch can result in induction, repression, or de-repression of other, regulated genes or cellular processes. Many such secondary regulatory effects are known and can be adapted for use with riboswitches. An advantage of riboswitches as the primary control for such regulation is that riboswitch 30 trigger molecules can be small, non-antigenic molecules.

Also disclosed are methods of identifying compounds that activate, deactivate or block a riboswitch. For examples, compounds that activate a riboswitch can be identified by bringing into contact a test compound and a riboswitch and assessing

activation of the riboswitch. If the riboswitch is activated, the test compound is identified as a compound that activates the riboswitch. Activation of a riboswitch can be assessed in any suitable manner. For example, the riboswitch can be linked to a reporter RNA and expression, expression level, or change in expression level of the reporter RNA can be measured in the presence and absence of the test compound. As another example, the riboswitch can include a conformation dependent label, the signal from which changes depending on the activation state of the riboswitch. Such a riboswitch preferably uses an aptamer domain from or derived from a naturally occurring riboswitch. As can be seen, assessment of activation of a riboswitch can be performed with the use of a control assay or measurement or without the use of a control assay or measurement. Methods for identifying compounds that deactivate a riboswitch can be performed in analogous ways.

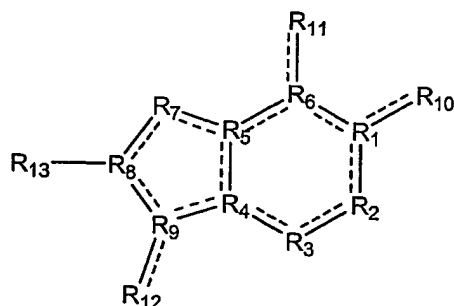
Identification of compounds that block a riboswitch can be accomplished in any suitable manner. For example, an assay can be performed for assessing activation or deactivation of a riboswitch in the presence of a compound known to activate or deactivate the riboswitch and in the presence of a test compound. If activation or deactivation is not observed as would be observed in the absence of the test compound, then the test compound is identified as a compound that blocks activation or deactivation of the riboswitch.

Also disclosed are compounds made by identifying a compound that activates, deactivates or blocks a riboswitch and manufacturing the identified compound. This can be accomplished by, for example, combining compound identification methods as disclosed elsewhere herein with methods for manufacturing the identified compounds. For example, compounds can be made by bringing into contact a test compound and a riboswitch, assessing activation of the riboswitch, and, if the riboswitch is activated by the test compound, manufacturing the test compound that activates the riboswitch as the compound.

Also disclosed are compounds made by checking activation, deactivation or blocking of a riboswitch by a compound and manufacturing the checked compound. This can be accomplished by, for example, combining compound activation, deactivation or blocking assessment methods as disclosed elsewhere herein with methods for manufacturing the checked compounds. For example, compounds can be made by bringing into contact a test compound and a riboswitch, assessing activation of the

riboswitch, and, if the riboswitch is activated by the test compound, manufacturing the test compound that activates the riboswitch as the compound. Checking compounds for their ability to activate, deactivate or block a riboswitch refers to both identification of compounds previously unknown to activate, deactivate or block a riboswitch and to
 5 assessing the ability of a compound to activate, deactivate or block a riboswitch where the compound was already known to activate, deactivate or block the riboswitch.

Specific compounds that can be used to activate riboswitches are also disclosed. Compounds useful with guanine-responsive riboswitches (and riboswitches derived from guanine-responsive riboswitches) include compounds having the formula

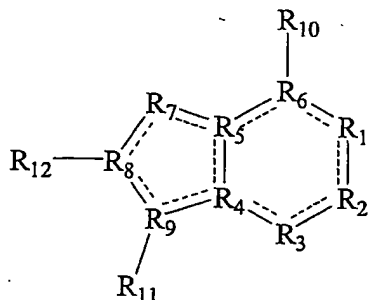


10

where the compound can bind a guanine-responsive riboswitch or derivative thereof, where, when the compound is bound to a guanine-responsive riboswitch or derivative, R_7 serves as a hydrogen bond acceptor, R_{10} serves as a hydrogen bond donor, R_{11} serves as a hydrogen bond acceptor, R_{12} serves as a hydrogen bond donor, where R_{13}
 15 is H, H_2 or is not present, where R_1 , R_2 , R_3 , R_4 , R_5 , R_6 , R_8 , and R_9 are each independently C, N, O, or S, and where ----- each independently represent a single or double bond.

Every compound within the above definition is intended to be and should be considered to be specifically disclosed herein. Further, every subgroup that can be identified within the above definition is intended to be and should be considered to be
 20 specifically disclosed herein. As a result, it is specifically contemplated that any compound, or subgroup of compounds can be either specifically included for or excluded from use or included in or excluded from a list of compounds. For example, as one option, a group of compounds is contemplated where each compound is as defined above but is not guanine, hypoxanthine, xanthine, or N^2 -methylguanine. As another example, a
 25 group of compounds is contemplated where each compound is as defined above and is able to activate a guanine-responsive riboswitch.

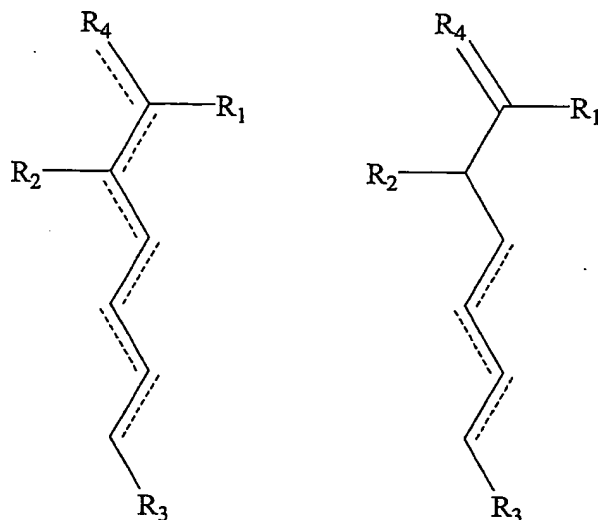
Compounds useful with adenine-responsive riboswitches (and riboswitches derived from adenine-responsive riboswitches) include compounds having the formula



where the compound can bind an adenine-responsive riboswitch or derivative thereof, where, when the compound is bound to an adenine-responsive riboswitch or derivative, R₁, R₃ and R₇ serve as hydrogen bond acceptors, and R₁₀ and R₁₁ serve as hydrogen bond donors, where R₁₂ is H, H₂ or is not present, where R₁, R₂, R₃, R₄, R₅, R₆, R₈, and R₉ are each independently C, N, O, or S, and where ----- each independently represent a single or double bond.

Every compound within the above definition is intended to be and should be considered to be specifically disclosed herein. Further, every subgroup that can be identified within the above definition is intended to be and should be considered to be specifically disclosed herein. As a result, it is specifically contemplated that any compound, or subgroup of compounds can be either specifically included for or excluded from use or included in or excluded from a list of compounds. For example, as one option, a group of compounds is contemplated where each compound is as defined above but is not adenine, 2,6-diaminopurine, or 2-amino purine. As another example, a group of compounds is contemplated where each compound is as defined above and is able to activate an adenine-responsive riboswitch.

Compounds useful with lysine-responsive riboswitches (and riboswitches derived from lysine-responsive riboswitches) include compounds having the formula

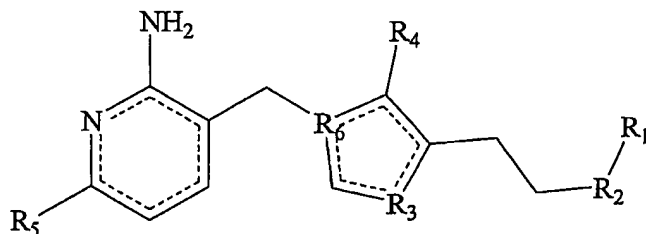


where the compound can bind a lysine-responsive riboswitch or derivative thereof, where R_2 and R_3 are each positively charged, where R_1 is negatively charged, where R_4 is C, N, O, or S, and where ----- each independently represent a single or double bond.

- 5 Also contemplated are compounds as defined above where R_2 and R_3 are each NH_3^+ and where R_1 is O^- .

Every compound within the above definition is intended to be and should be considered to be specifically disclosed herein. Further, every subgroup that can be identified within the above definition is intended to be and should be considered to be specifically disclosed herein. As a result, it is specifically contemplated that any compound, or subgroup of compounds can be either specifically included for or excluded from use or included in or excluded from a list of compounds. For example, as one option, a group of compounds is contemplated where each compound is as defined above but is not lysine. As another example, a group of compounds is contemplated where each compound is as defined above and is able to activate a lysine-responsive riboswitch.

Compounds useful with TPP-responsive riboswitches (and riboswitches derived from lysine-responsive riboswitches) include compounds having the formula



where the compound can bind a TPP-responsive riboswitch or derivative thereof, where R_1 is positively charged, where R_2 and R_3 are each independently C, O, or S, where R_4 is CH_3 , NH_2 , OH, SH, H or not present, where R_5 is CH_3 , NH_2 , OH, SH, or H, where R_6 is C or N, and where ----- each independently represent a single or double bond. Also
5 contemplated are compounds as defined above where R_1 is phosphate, diphosphate or triphosphate.

Every compound within the above definition is intended to be and should be considered to be specifically disclosed herein. Further, every subgroup that can be identified within the above definition is intended to be and should be considered to be
10 specifically disclosed herein. As a result, it is specifically contemplated that any compound, or subgroup of compounds can be either specifically included for or excluded from use or included in or excluded from a list of compounds. For example, as one option, a group of compounds is contemplated where each compound is as defined above but is not TPP, TP or thiamine. As another example, a group of compounds is
15 contemplated where each compound is as defined above and is able to activate a TPP-responsive riboswitch.

D. Constructs, Vectors and Expression Systems

The disclosed riboswitches can be used in with any suitable expression system. Recombinant expression is usefully accomplished using a vector, such as a plasmid. The
20 vector can include a promoter operably linked to riboswitch-encoding sequence and RNA to be expression (e.g., RNA encoding a protein). The vector can also include other elements required for transcription and translation. As used herein, vector refers to any carrier containing exogenous DNA. Thus, vectors are agents that transport the exogenous nucleic acid into a cell without degradation and include a promoter yielding
25 expression of the nucleic acid in the cells into which it is delivered. Vectors include but are not limited to plasmids, viral nucleic acids, viruses, phage nucleic acids, phages, cosmids, and artificial chromosomes. A variety of prokaryotic and eukaryotic expression vectors suitable for carrying riboswitch-regulated constructs can be produced. Such expression vectors include, for example, pET, pET3d, pCR2.1, pBAD, pUC, and yeast
30 vectors. The vectors can be used, for example, in a variety of *in vivo* and *in vitro* situation.

Viral vectors include adenovirus, adeno-associated virus, herpes virus, vaccinia virus, polio virus, AIDS virus, neuronal trophic virus, Sindbis and other RNA viruses,

including these viruses with the HIV backbone. Also useful are any viral families which share the properties of these viruses which make them suitable for use as vectors.

Retroviral vectors, which are described in Verma (1985), include Murine Maloney Leukemia virus, MMLV, and retroviruses that express the desirable properties of MMLV as a vector. Typically, viral vectors contain, nonstructural early genes, structural late genes, an RNA polymerase III transcript, inverted terminal repeats necessary for replication and encapsidation, and promoters to control the transcription and replication of the viral genome. When engineered as vectors, viruses typically have one or more of the early genes removed and a gene or gene/promotor cassette is inserted into the viral genome in place of the removed viral DNA.

A "promoter" is generally a sequence or sequences of DNA that function when in a relatively fixed location in regard to the transcription start site. A "promoter" contains core elements required for basic interaction of RNA polymerase and transcription factors and can contain upstream elements and response elements.

"Enhancer" generally refers to a sequence of DNA that functions at no fixed distance from the transcription start site and can be either 5' (Laimins, 1981) or 3' (Lusky et al., 1983) to the transcription unit. Furthermore, enhancers can be within an intron (Banerji et al., 1983) as well as within the coding sequence itself (Osborne et al., 1984). They are usually between 10 and 300 bp in length, and they function in cis.

Enhancers function to increase transcription from nearby promoters. Enhancers, like promoters, also often contain response elements that mediate the regulation of transcription. Enhancers often determine the regulation of expression.

Expression vectors used in eukaryotic host cells (yeast, fungi, insect, plant, animal, human or nucleated cells) can also contain sequences necessary for the termination of transcription which can affect mRNA expression. These regions are transcribed as polyadenylated segments in the untranslated portion of the mRNA encoding tissue factor protein. The 3' untranslated regions also include transcription termination sites. It is preferred that the transcription unit also contain a polyadenylation region. One benefit of this region is that it increases the likelihood that the transcribed unit will be processed and transported like mRNA. The identification and use of polyadenylation signals in expression constructs is well established. It is preferred that homologous polyadenylation signals be used in the transgene constructs.

The vector can include nucleic acid sequence encoding a marker product. This marker product is used to determine if the gene has been delivered to the cell and once delivered is being expressed. Preferred marker genes are the *E. Coli* lacZ gene which encodes β -galactosidase and green fluorescent protein.

5 In some embodiments the marker can be a selectable marker. When such selectable markers are successfully transferred into a host cell, the transformed host cell can survive if placed under selective pressure. There are two widely used distinct categories of selective regimes. The first category is based on a cell's metabolism and the use of a mutant cell line which lacks the ability to grow independent of a
10 supplemented media. The second category is dominant selection which refers to a selection scheme used in any cell type and does not require the use of a mutant cell line. These schemes typically use a drug to arrest growth of a host cell. Those cells which have a novel gene would express a protein conveying drug resistance and would survive the selection. Examples of such dominant selection use the drugs neomycin, (Southern
15 and Berg, 1982), mycophenolic acid, (Mulligan and Berg, 1980) or hygromycin (Sugden et al., 1985).

Gene transfer can be obtained using direct transfer of genetic material, in but not limited to, plasmids, viral vectors, viral nucleic acids, phage nucleic acids, phages, cosmids, and artificial chromosomes, or via transfer of genetic material in cells or
20 carriers such as cationic liposomes. Such methods are well known in the art and readily adaptable for use in the method described herein. Transfer vectors can be any nucleotide construction used to deliver genes into cells (e.g., a plasmid), or as part of a general strategy to deliver genes, e.g., as part of recombinant retrovirus or adenovirus (Ram et al. Cancer Res. 53:83-88, (1993)). Appropriate means for transfection, including viral
25 vectors, chemical transfectants, or physico-mechanical methods such as electroporation and direct diffusion of DNA, are described by, for example, Wolff, J. A., et al., Science, 247, 1465-1468, (1990); and Wolff, J. A. Nature, 352, 815-818, (1991).

1. Viral Vectors

Preferred viral vectors are Adenovirus, Adeno-associated virus, Herpes virus,
30 Vaccinia virus, Polio virus, AIDS virus, neuronal trophic virus, Sindbis and other RNA viruses, including these viruses with the HIV backbone. Also preferred are any viral families which share the properties of these viruses which make them suitable for use as vectors. Preferred retroviruses include Murine Maloney Leukemia virus, MMLV, and

retroviruses that express the desirable properties of MMLV as a vector. Retroviral vectors are able to carry a larger genetic payload, i.e., a transgene or marker gene, than other viral vectors, and for this reason are a commonly used vector. However, they are not useful in non-proliferating cells. Adenovirus vectors are relatively stable and easy to work with, have high titers, and can be delivered in aerosol formulation, and can transfect non-dividing cells. Pox viral vectors are large and have several sites for inserting genes, they are thermostable and can be stored at room temperature. A preferred embodiment is a viral vector which has been engineered so as to suppress the immune response of the host organism, elicited by the viral antigens. Preferred vectors of this type will carry coding regions for Interleukin 8 or 10.

Viral vectors have higher transaction (ability to introduce genes) abilities than do most chemical or physical methods to introduce genes into cells. Typically, viral vectors contain, nonstructural early genes, structural late genes, an RNA polymerase III transcript, inverted terminal repeats necessary for replication and encapsidation, and promoters to control the transcription and replication of the viral genome. When engineered as vectors, viruses typically have one or more of the early genes removed and a gene or gene/promotor cassette is inserted into the viral genome in place of the removed viral DNA. Constructs of this type can carry up to about 8 kb of foreign genetic material. The necessary functions of the removed early genes are typically supplied by cell lines which have been engineered to express the gene products of the early genes in trans.

i. Retroviral Vectors

A retrovirus is an animal virus belonging to the virus family of Retroviridae, including any types, subfamilies, genus, or tropisms. Retroviral vectors, in general, are described by Verma, I.M., Retroviral vectors for gene transfer. In Microbiology-1985, American Society for Microbiology, pp. 229-232, Washington, (1985), which is incorporated by reference herein. Examples of methods for using retroviral vectors for gene therapy are described in U.S. Patent Nos. 4,868,116 and 4,980,286; PCT applications WO 90/02806 and WO 89/07136; and Mulligan, (Science 260:926-932 (1993)); the teachings of which are incorporated herein by reference.

A retrovirus is essentially a package which has packed into it nucleic acid cargo. The nucleic acid cargo carries with it a packaging signal, which ensures that the replicated daughter molecules will be efficiently packaged within the package coat. In

addition to the package signal, there are a number of molecules which are needed in cis, for the replication, and packaging of the replicated virus. Typically a retroviral genome, contains the gag, pol, and env genes which are involved in the making of the protein coat. It is the gag, pol, and env genes which are typically replaced by the foreign DNA that it is to be transferred to the target cell. Retrovirus vectors typically contain a packaging signal for incorporation into the package coat, a sequence which signals the start of the gag transcription unit, elements necessary for reverse transcription, including a primer binding site to bind the tRNA primer of reverse transcription, terminal repeat sequences that guide the switch of RNA strands during DNA synthesis, a purine rich sequence 5' to the 3' LTR that serve as the priming site for the synthesis of the second strand of DNA synthesis, and specific sequences near the ends of the LTRs that enable the insertion of the DNA state of the retrovirus to insert into the host genome. The removal of the gag, pol, and env genes allows for about 8 kb of foreign sequence to be inserted into the viral genome, become reverse transcribed, and upon replication be packaged into a new retroviral particle. This amount of nucleic acid is sufficient for the delivery of a one to many genes depending on the size of each transcript. It is preferable to include either positive or negative selectable markers along with other genes in the insert.

Since the replication machinery and packaging proteins in most retroviral vectors have been removed (gag, pol, and env), the vectors are typically generated by placing them into a packaging cell line. A packaging cell line is a cell line which has been transfected or transformed with a retrovirus that contains the replication and packaging machinery, but lacks any packaging signal. When the vector carrying the DNA of choice is transfected into these cell lines, the vector containing the gene of interest is replicated and packaged into new retroviral particles, by the machinery provided in cis by the helper cell. The genomes for the machinery are not packaged because they lack the necessary signals.

ii. Adenoviral Vectors

The construction of replication-defective adenoviruses has been described (Berkner et al., J. Virology 61:1213-1220 (1987); Massie et al., Mol. Cell. Biol. 6:2872-2883 (1986); Haj-Ahmad et al., J. Virology 57:267-274 (1986); Davidson et al., J. Virology 61:1226-1239 (1987); Zhang "Generation and identification of recombinant adenovirus by liposome-mediated transfection and PCR analysis" BioTechniques

15:868-872 (1993)). The benefit of the use of these viruses as vectors is that they are limited in the extent to which they can spread to other cell types, since they can replicate within an initial infected cell, but are unable to form new infectious viral particles. Recombinant adenoviruses have been shown to achieve high efficiency gene transfer after direct, in vivo delivery to airway epithelium, hepatocytes, vascular endothelium, CNS parenchyma and a number of other tissue sites (Morsy, *J. Clin. Invest.* 92:1580-1586 (1993); Kirshenbaum, *J. Clin. Invest.* 92:381-387 (1993); Roessler, *J. Clin. Invest.* 92:1085-1092 (1993); Moullier, *Nature Genetics* 4:154-159 (1993); La Salle, *Science* 259:988-990 (1993); Gomez-Foix, *J. Biol. Chem.* 267:25129-25134 (1992); Rich, *Human Gene Therapy* 4:461-476 (1993); Zabner, *Nature Genetics* 6:75-83 (1994); Guzman, *Circulation Research* 73:1201-1207 (1993); Bout, *Human Gene Therapy* 5:3-10 (1994); Zabner, *Cell* 75:207-216 (1993); Caillaud, *Eur. J. Neuroscience* 5:1287-1291 (1993); and Ragot, *J. Gen. Virology* 74:501-507 (1993)). Recombinant adenoviruses achieve gene transduction by binding to specific cell surface receptors, after which the virus is internalized by receptor-mediated endocytosis, in the same manner as wild type or replication-defective adenovirus (Chardonnet and Dales, *Virology* 40:462-477 (1970); Brown and Burlingham, *J. Virology* 12:386-396 (1973); Svensson and Persson, *J. Virology* 55:442-449 (1985); Seth, et al., *J. Virol.* 51:650-655 (1984); Seth, et al., *Mol. Cell. Biol.* 4:1528-1533 (1984); Varga et al., *J. Virology* 65:6061-6070 (1991); Wickham et al., *Cell* 73:309-319 (1993)).

A preferred viral vector is one based on an adenovirus which has had the E1 gene removed and these virions are generated in a cell line such as the human 293 cell line. In another preferred embodiment both the E1 and E3 genes are removed from the adenovirus genome.

Another type of viral vector is based on an adeno-associated virus (AAV). This defective parvovirus is a preferred vector because it can infect many cell types and is nonpathogenic to humans. AAV type vectors can transport about 4 to 5 kb and wild type AAV is known to stably insert into chromosome 19. Vectors which contain this site specific integration property are preferred. An especially preferred embodiment of this type of vector is the P4.1 C vector produced by Avigen, San Francisco, CA, which can contain the herpes simplex virus thymidine kinase gene, HSV-tk, and/or a marker gene, such as the gene encoding the green fluorescent protein, GFP.

The inserted genes in viral and retroviral usually contain promoters, and/or enhancers to help control the expression of the desired gene product. A promoter is generally a sequence or sequences of DNA that function when in a relatively fixed location in regard to the transcription start site. A promoter contains core elements required for basic interaction of RNA polymerase and transcription factors, and can contain upstream elements and response elements.

2. Viral Promoters and Enhancers

Preferred promoters controlling transcription from vectors in mammalian host cells can be obtained from various sources, for example, the genomes of viruses such as: polyoma, Simian Virus 40 (SV40), adenovirus, retroviruses, hepatitis-B virus and most preferably cytomegalovirus, or from heterologous mammalian promoters, e.g. beta actin promoter. The early and late promoters of the SV40 virus are conveniently obtained as an SV40 restriction fragment which also contains the SV40 viral origin of replication (Fiers et al., *Nature*, 273: 113 (1978)). The immediate early promoter of the human cytomegalovirus is conveniently obtained as a HindIII E restriction fragment (Greenway, P.J. et al., *Gene* 18: 355-360 (1982)). Of course, promoters from the host cell or related species also are useful herein.

Enhancer generally refers to a sequence of DNA that functions at no fixed distance from the transcription start site and can be either 5' (Laimins, L. et al., *Proc. Natl. Acad. Sci.* 78: 993 (1981)) or 3' (Lusky, M.L., et al., *Mol. Cell Bio.* 3: 1108 (1983)) to the transcription unit. Furthermore, enhancers can be within an intron (Banerji, J.L. et al., *Cell* 33: 729 (1983)) as well as within the coding sequence itself (Osborne, T.F., et al., *Mol. Cell Bio.* 4: 1293 (1984)). They are usually between 10 and 300 bp in length, and they function in cis. Enhancers function to increase transcription from nearby promoters. Enhancers also often contain response elements that mediate the regulation of transcription. Promoters can also contain response elements that mediate the regulation of transcription. Enhancers often determine the regulation of expression of a gene. While many enhancer sequences are now known from mammalian genes (globin, elastase, albumin, α -fetoprotein and insulin), typically one will use an enhancer from a eukaryotic cell virus. Preferred examples are the SV40 enhancer on the late side of the replication origin (bp 100-270), the cytomegalovirus early promoter enhancer, the polyoma enhancer on the late side of the replication origin, and adenovirus enhancers.

The promotor and/or enhancer can be specifically activated either by light or specific chemical events which trigger their function. Systems can be regulated by reagents such as tetracycline and dexamethasone. There are also ways to enhance viral vector gene expression by exposure to irradiation, such as gamma irradiation, or
5 alkylating chemotherapy drugs.

It is preferred that the promoter and/or enhancer region be active in all eukaryotic cell types. A preferred promoter of this type is the CMV promoter (650 bases). Other preferred promoters are SV40 promoters, cytomegalovirus (full length promoter), and retroviral vector LTF.

10 It has been shown that all specific regulatory elements can be cloned and used to construct expression vectors that are selectively expressed in specific cell types such as melanoma cells. The glial fibrillary acetic protein (GFAP) promoter has been used to selectively express genes in cells of glial origin.

Expression vectors used in eukaryotic host cells (yeast, fungi, insect, plant,
15 animal, human or nucleated cells) can also contain sequences necessary for the termination of transcription which can affect mRNA expression. These regions are transcribed as polyadenylated segments in the untranslated portion of the mRNA encoding tissue factor protein. The 3' untranslated regions also include transcription termination sites. It is preferred that the transcription unit also contain a
20 polyadenylation region. One benefit of this region is that it increases the likelihood that the transcribed unit will be processed and transported like mRNA. The identification and use of polyadenylation signals in expression constructs is well established. It is preferred that homologous polyadenylation signals be used in the transgene constructs. In a preferred embodiment of the transcription unit, the polyadenylation region is derived
25 from the SV40 early polyadenylation signal and consists of about 400 bases. It is also preferred that the transcribed units contain other standard sequences alone or in combination with the above sequences improve expression from, or stability of, the construct.

3. Markers

30 The vectors can include nucleic acid sequence encoding a marker product. This marker product is used to determine if the gene has been delivered to the cell and once

delivered is being expressed. Preferred marker genes are the E. Coli lacZ gene which encodes β -galactosidase and green fluorescent protein.

In some embodiments the marker can be a selectable marker. Examples of suitable selectable markers for mammalian cells are dihydrofolate reductase (DHFR), thymidine kinase, neomycin, neomycin analog G418, hygromycin, and puromycin. When such selectable markers are successfully transferred into a mammalian host cell, the transformed mammalian host cell can survive if placed under selective pressure. There are two widely used distinct categories of selective regimes. The first category is based on a cell's metabolism and the use of a mutant cell line which lacks the ability to grow independent of a supplemented media. Two examples are: CHO DHFR⁻ cells and mouse LTK⁻ cells. These cells lack the ability to grow without the addition of such nutrients as thymidine or hypoxanthine. Because these cells lack certain genes necessary for a complete nucleotide synthesis pathway, they cannot survive unless the missing nucleotides are provided in a supplemented media. An alternative to supplementing the media is to introduce an intact DHFR or TK gene into cells lacking the respective genes, thus altering their growth requirements. Individual cells which were not transformed with the DHFR or TK gene will not be capable of survival in non-supplemented media.

The second category is dominant selection which refers to a selection scheme used in any cell type and does not require the use of a mutant cell line. These schemes typically use a drug to arrest growth of a host cell. Those cells which would express a protein conveying drug resistance and would survive the selection. Examples of such dominant selection use the drugs neomycin, (Southern P. and Berg, P., J. Molec. Appl. Genet. 1: 327 (1982)), mycophenolic acid, (Mulligan, R.C. and Berg, P. Science 209: 1422 (1980)) or hygromycin, (Sugden, B. et al., Mol. Cell. Biol. 5: 410-413 (1985)). The three examples employ bacterial genes under eukaryotic control to convey resistance to the appropriate drug G418 or neomycin (geneticin), xgpt (mycophenolic acid) or hygromycin, respectively. Others include the neomycin analog G418 and puramycin.

E. Biosensor Riboswitches

Also disclosed are biosensor riboswitches. Biosensor riboswitches are engineered riboswitches that produce a detectable signal in the presence of their cognate trigger molecule. Useful biosensor riboswitches can be triggered at or above threshold levels of the trigger molecules. Biosensor riboswitches can be designed for use in vivo

or in vitro. For example, biosensor riboswitches operably linked to a reporter RNA that encodes a protein that serves as or is involved in producing a signal can be used in vivo by engineering a cell or organism to harbor a nucleic acid construct encoding the riboswitch/reporter RNA. An example of a biosensor riboswitch for use in vitro is a
5 riboswitch that includes a conformation dependent label, the signal from which changes depending on the activation state of the riboswitch. Such a biosensor riboswitch preferably uses an aptamer domain from or derived from a naturally occurring riboswitch.

F. Reporter Proteins and Peptides

10 For assessing activation of a riboswitch, or for biosensor riboswitches, a reporter protein or peptide can be used. The reporter protein or peptide can be encoded by the RNA the expression of which is regulated by the riboswitch. The examples describe the use of some specific reporter proteins. The use of reporter proteins and peptides is well known and can be adapted easily for use with riboswitches. The reporter proteins can be
15 any protein or peptide that can be detected or that produces a detectable signal. Preferably, the presence of the protein or peptide can be detected using standard techniques (e.g., radioimmunoassay, radio-labeling, immunoassay, assay for enzymatic activity, absorbance, fluorescence, luminescence, and Western blot). More preferably, the level of the reporter protein is easily quantifiable using standard techniques even at
20 low levels. Useful reporter proteins include luciferases, green fluorescent proteins and their derivatives, such as firefly luciferase (FL) from *Photinus pyralis*, and Renilla luciferase (RL) from *Renilla reniformis*.

G. Conformation Dependent Labels

Conformation dependent labels refer to all labels that produce a change in
25 fluorescence intensity or wavelength based on a change in the form or conformation of the molecule or compound (such as a riboswitch) with which the label is associated. Examples of conformation dependent labels used in the context of probes and primers include molecular beacons, Amplifluors, FRET probes, cleavable FRET probes, TaqMan probes, scorpion primers, fluorescent triplex oligos including but not limited to triplex
30 molecular beacons or triplex FRET probes, fluorescent water-soluble conjugated polymers, PNA probes and QPNA probes. Such labels, and, in particular, the principles of their function, can be adapted for use with riboswitches. Several types of

conformation dependent labels are reviewed in Schweitzer and Kingsmore, Curr. Opin. Biotech. 12:21-27 (2001).

Stem quenched labels, a form of conformation dependent labels, are fluorescent labels positioned on a nucleic acid such that when a stem structure forms a quenching moiety is brought into proximity such that fluorescence from the label is quenched. When the stem is disrupted (such as when a riboswitch containing the label is activated), the quenching moiety is no longer in proximity to the fluorescent label and fluorescence increases. Examples of this effect can be found in molecular beacons, fluorescent triplex oligos, triplex molecular beacons, triplex FRET probes, and QPNA probes, the operational principles of which can be adapted for use with riboswitches.

Stem activated labels, a form of conformation dependent labels, are labels or pairs of labels where fluorescence is increased or altered by formation of a stem structure. Stem activated labels can include an acceptor fluorescent label and a donor moiety such that, when the acceptor and donor are in proximity (when the nucleic acid strands containing the labels form a stem structure), fluorescence resonance energy transfer from the donor to the acceptor causes the acceptor to fluoresce. Stem activated labels are typically pairs of labels positioned on nucleic acid molecules (such as riboswitches) such that the acceptor and donor are brought into proximity when a stem structure is formed in the nucleic acid molecule. If the donor moiety of a stem activated label is itself a fluorescent label, it can release energy as fluorescence (typically at a different wavelength than the fluorescence of the acceptor) when not in proximity to an acceptor (that is, when a stem structure is not formed). When the stem structure forms, the overall effect would then be a reduction of donor fluorescence and an increase in acceptor fluorescence. FRET probes are an example of the use of stem activated labels, the operational principles of which can be adapted for use with riboswitches.

H. Detection Labels

To aid in detection and quantitation of riboswitch activation, deactivation or blocking, or expression of nucleic acids or protein produced upon activation, deactivation or blocking of riboswitches, detection labels can be incorporated into detection probes or detection molecules or directly incorporated into expressed nucleic acids or proteins. As used herein, a detection label is any molecule that can be associated with nucleic acid or protein, directly or indirectly, and which results in a measurable, detectable signal, either directly or indirectly. Many such labels are known to those of

skill in the art. Examples of detection labels suitable for use in the disclosed method are radioactive isotopes, fluorescent molecules, phosphorescent molecules, enzymes, antibodies, and ligands.

Examples of suitable fluorescent labels include fluorescein isothiocyanate (FITC), 5,6-carboxymethyl fluorescein, Texas red, nitrobenz-2-oxa-1,3-diazol-4-yl (NBD), coumarin, dansyl chloride, rhodamine, amino-methyl coumarin (AMCA), Eosin, Erythrosin, BODIPY[®], Cascade Blue[®], Oregon Green[®], pyrene, lissamine, xanthenes, acridines, oxazines, phycoerythrin, macrocyclic chelates of lanthanide ions such as quantum dyeTM, fluorescent energy transfer dyes, such as thiazole orange-ethidium heterodimer, and the cyanine dyes Cy3, Cy3.5, Cy5, Cy5.5 and Cy7. Examples of other specific fluorescent labels include 3-Hydroxypyrene 5,8,10-Tri Sulfonic acid, 5-Hydroxy Tryptamine (5-HT), Acid Fuchsin, Alizarin Complexon, Alizarin Red, Allophycocyanin, Aminocoumarin, Anthroyl Stearate, Astrazon Brilliant Red 4G, Astrazon Orange R, Astrazon Red 6B, Astrazon Yellow 7 GLL, Atabrine, Auramine, Aurophosphine, Aurophosphine G, BAO 9 (Bisaminophenyloxadiazole), BCECF, Berberine Sulphate, Bisbenzamide, Blancophor FFG Solution, Blancophor SV, Bodipy F1, Brilliant Sulphoflavin FF, Calcion Blue, Calcium Green, Calcofluor RW Solution, Calcofluor White, Calcophor White ABT Solution, Calcophor White Standard Solution, Carbostryl, Cascade Yellow, Catecholamine, Chinacrine, Coriphosphine O, Coumarin-Phalloidin, CY3.1 8, CY5.1 8, CY7, Dans (1-Dimethyl Amino Naphaline 5 Sulphonic Acid), Dansa (Diamino Naphtyl Sulphonic Acid), Dansyl NH-CH₃, Diamino Phenyl Oxydiazole (DAO), Dimethylamino-5-Sulphonic acid, Dipyrrometheneboron Difluoride, Diphenyl Brilliant Flavine 7GFF, Dopamine, Erythrosin ITC, Euchrysin, FIF (Formaldehyde Induced Fluorescence), Flazo Orange, Fluo 3, Fluorescamine, Fura-2, Genacryl Brilliant Red B, Genacryl Brilliant Yellow 10GF, Genacryl Pink 3G, Genacryl Yellow 5GF, Gloxalic Acid, Granular Blue, Haematoporphyrin, Indo-1, Intrawhite Cf Liquid, Leucophor PAF, Leucophor SF, Leucophor WS, Lissamine Rhodamine B200 (RD200), Lucifer Yellow CH, Lucifer Yellow VS, Magdala Red, Marina Blue, Maxilon Brilliant Flavin 10 GFF, Maxilon Brilliant Flavin 8 GFF, MPS (Methyl Green Pyronine Stilbene), Mithramycin, NBD Amine, Nitrobenzoxadidole, Noradrenaline, Nuclear Fast Red, Nuclear Yellow, Nylosan Brilliant Flavin E8G, Oxadiazole, Pacific Blue, Pararosanine (Feulgen), Phorwite AR Solution, Phorwite BKL, Phorwite Rev, Phorwite RPA, Phosphine 3R, Phthalocyanine, Phycoerythrin R, Polyazaindacene Pontochrome

Blue Black, Porphyrin, Primuline, Procion Yellow, Pyronine, Pyronine B, Pyrozal
Brilliant Flavin 7GF, Quinacrine Mustard, Rhodamine 123, Rhodamine 5 GLD,
Rhodamine 6G, Rhodamine B, Rhodamine B 200, Rhodamine B Extra, Rhodamine BB,
Rhodamine BG, Rhodamine WT, Serotonin, Sevron Brilliant Red 2B, Sevron Brilliant
5 Red 4G, Sevron Brilliant Red B, Sevron Orange, Sevron Yellow L, SITS (Primuline),
SITS (Stilbene Isothiosulphonic acid), Stilbene, Snarf 1, sulpho Rhodamine B Can C,
Sulpho Rhodamine G Extra, Tetracycline, Thiazine Red R, Thioflavin S, Thioflavin
TCN, Thioflavin 5, Thiolyte, Thiozol Orange, Tinopol CBS, True Blue, Ultralite,
Uranine B, Uvitex SFC, Xylene Orange, and XRITC.

10 Useful fluorescent labels are fluorescein (5-carboxyfluorescein-N-
hydroxysuccinimide ester), rhodamine (5,6-tetramethyl rhodamine), and the cyanine
dyes Cy3, Cy3.5, Cy5, Cy5.5 and Cy7. The absorption and emission maxima,
respectively, for these fluors are: FITC (490 nm; 520 nm), Cy3 (554 nm; 568 nm), Cy3.5
(581 nm; 588 nm), Cy5 (652 nm; 672 nm), Cy5.5 (682 nm; 703 nm) and Cy7 (755 nm;
15 778 nm), thus allowing their simultaneous detection. Other examples of fluorescein dyes
include 6-carboxyfluorescein (6-FAM), 2',4',1,4,-tetrachlorofluorescein (TET),
2',4',5',7',1,4-hexachlorofluorescein (HEX), 2',7'-dimethoxy-4', 5'-dichloro-6-
carboxyrhodamine (JOE), 2'-chloro-5'-fluoro-7',8'-fused phenyl-1,4-dichloro-6-
carboxyfluorescein (NED), and 2'-chloro-7'-phenyl-1,4-dichloro-6-carboxyfluorescein
20 (VIC). Fluorescent labels can be obtained from a variety of commercial sources,
including Amersham Pharmacia Biotech, Piscataway, NJ; Molecular Probes, Eugene,
OR; and Research Organics, Cleveland, Ohio.

Additional labels of interest include those that provide for signal only when the
probe with which they are associated is specifically bound to a target molecule, where
25 such labels include: "molecular beacons" as described in Tyagi & Kramer, *Nature*
Biotechnology (1996) 14:303 and EP 0 070 685 B1. Other labels of interest include
those described in U.S. Pat. No. 5,563,037; WO 97/17471 and WO 97/17076.

Labeled nucleotides are a useful form of detection label for direct incorporation
into expressed nucleic acids during synthesis. Examples of detection labels that can be
30 incorporated into nucleic acids include nucleotide analogs such as BrdUrd (5-
bromodeoxyuridine, Hoy and Schimke, *Mutation Research* 290:217-230 (1993)),
aminoallyldeoxyuridine (Henegariu *et al.*, *Nature Biotechnology* 18:345-348 (2000)), 5-
methylcytosine (Sano *et al.*, *Biochim. Biophys. Acta* 951:157-165 (1988)), bromouridine

(Wansick *et al.*, *J. Cell Biology* 122:283-293 (1993)) and nucleotides modified with biotin (Langer *et al.*, *Proc. Natl. Acad. Sci. USA* 78:6633 (1981)) or with suitable haptens such as digoxigenin (Kerkhof, *Anal. Biochem.* 205:359-364 (1992)). Suitable fluorescence-labeled nucleotides are Fluorescein-isothiocyanate-dUTP, Cyanine-3-dUTP and Cyanine-5-dUTP (Yu *et al.*, *Nucleic Acids Res.*, 22:3226-3232 (1994)). A preferred nucleotide analog detection label for DNA is BrdUrd (bromodeoxyuridine, BrdUrd, BrdU, BUdR, Sigma-Aldrich Co). Other useful nucleotide analogs for incorporation of detection label into DNA are AA-dUTP (aminoallyl-deoxyuridine triphosphate, Sigma-Aldrich Co.), and 5-methyl-dCTP (Roche Molecular Biochemicals). A useful nucleotide analog for incorporation of detection label into RNA is biotin-16-UTP (biotin-16-uridine-5'-triphosphate, Roche Molecular Biochemicals). Fluorescein, Cy3, and Cy5 can be linked to dUTP for direct labelling. Cy3.5 and Cy7 are available as avidin or anti-digoxigenin conjugates for secondary detection of biotin- or digoxigenin-labelled probes.

Detection labels that are incorporated into nucleic acid, such as biotin, can be subsequently detected using sensitive methods well-known in the art. For example, biotin can be detected using streptavidin-alkaline phosphatase conjugate (Tropix, Inc.), which is bound to the biotin and subsequently detected by chemiluminescence of suitable substrates (for example, chemiluminescent substrate CSPD: disodium, 3-(4-methoxyspiro-[1,2,-dioxetane-3-2'-(5'-chloro)tricyclo [3.3.1.1^{3,7}]decane]-4-yl) phenyl phosphate; Tropix, Inc.). Labels can also be enzymes, such as alkaline phosphatase, soybean peroxidase, horseradish peroxidase and polymerases, that can be detected, for example, with chemical signal amplification or by using a substrate to the enzyme which produces light (for example, a chemiluminescent 1,2-dioxetane substrate) or fluorescent signal.

Molecules that combine two or more of these detection labels are also considered detection labels. Any of the known detection labels can be used with the disclosed probes, tags, molecules and methods to label and detect activated or deactivated riboswitches or nucleic acid or protein produced in the disclosed methods. Methods for detecting and measuring signals generated by detection labels are also known to those of skill in the art. For example, radioactive isotopes can be detected by scintillation counting or direct visualization; fluorescent molecules can be detected with fluorescent spectrophotometers; phosphorescent molecules can be detected with a spectrophotometer

or directly visualized with a camera; enzymes can be detected by detection or visualization of the product of a reaction catalyzed by the enzyme; antibodies can be detected by detecting a secondary detection label coupled to the antibody. As used herein, detection molecules are molecules which interact with a compound or composition to be detected and to which one or more detection labels are coupled.

I. Sequence Similarities

It is understood that as discussed herein the use of the terms homology and identity mean the same thing as similarity. Thus, for example, if the use of the word homology is used between two sequences (non-natural sequences, for example) it is understood that this is not necessarily indicating an evolutionary relationship between these two sequences, but rather is looking at the similarity or relatedness between their nucleic acid sequences. Many of the methods for determining homology between two evolutionarily related molecules are routinely applied to any two or more nucleic acids or proteins for the purpose of measuring sequence similarity regardless of whether they are evolutionarily related or not.

In general, it is understood that one way to define any known variants and derivatives or those that might arise, of the disclosed riboswitches, aptamers, expression platforms, genes and proteins herein, is through defining the variants and derivatives in terms of homology to specific known sequences. This identity of particular sequences disclosed herein is also discussed elsewhere herein. In general, variants of riboswitches, aptamers, expression platforms, genes and proteins herein disclosed typically have at least, about 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, or 99 percent homology to a stated sequence or a native sequence. Those of skill in the art readily understand how to determine the homology of two proteins or nucleic acids, such as genes. For example, the homology can be calculated after aligning the two sequences so that the homology is at its highest level.

Another way of calculating homology can be performed by published algorithms. Optimal alignment of sequences for comparison can be conducted by the local homology algorithm of Smith and Waterman Adv. Appl. Math. 2: 482 (1981), by the homology alignment algorithm of Needleman and Wunsch, J. Mol Biol. 48: 443 (1970), by the search for similarity method of Pearson and Lipman, Proc. Natl. Acad. Sci. U.S.A. 85: 2444 (1988), by computerized implementations of these algorithms (GAP, BESTFIT,

FASTA, and TFASTA in the Wisconsin Genetics Software Package, Genetics Computer Group, 575 Science Dr., Madison, WI), or by inspection.

The same types of homology can be obtained for nucleic acids by for example the algorithms disclosed in Zuker, M. *Science* 244:48-52, 1989, Jaeger et al. *Proc. Natl. Acad. Sci. USA* 86:7706-7710, 1989, Jaeger et al. *Methods Enzymol.* 183:281-306, 1989 which are herein incorporated by reference for at least material related to nucleic acid alignment. It is understood that any of the methods typically can be used and that in certain instances the results of these various methods can differ, but the skilled artisan understands if identity is found with at least one of these methods, the sequences would be said to have the stated identity.

For example, as used herein, a sequence recited as having a particular percent homology to another sequence refers to sequences that have the recited homology as calculated by any one or more of the calculation methods described above. For example, a first sequence has 80 percent homology, as defined herein, to a second sequence if the first sequence is calculated to have 80 percent homology to the second sequence using the Zuker calculation method even if the first sequence does not have 80 percent homology to the second sequence as calculated by any of the other calculation methods. As another example, a first sequence has 80 percent homology, as defined herein, to a second sequence if the first sequence is calculated to have 80 percent homology to the second sequence using both the Zuker calculation method and the Pearson and Lipman calculation method even if the first sequence does not have 80 percent homology to the second sequence as calculated by the Smith and Waterman calculation method, the Needleman and Wunsch calculation method, the Jaeger calculation methods, or any of the other calculation methods. As yet another example, a first sequence has 80 percent homology, as defined herein, to a second sequence if the first sequence is calculated to have 80 percent homology to the second sequence using each of calculation methods (although, in practice, the different calculation methods will often result in different calculated homology percentages).

J. Hybridization and Selective Hybridization

The term hybridization typically means a sequence driven interaction between at least two nucleic acid molecules, such as a primer or a probe and a riboswitch or a gene. Sequence driven interaction means an interaction that occurs between two nucleotides or nucleotide analogs or nucleotide derivatives in a nucleotide specific manner. For

example, G interacting with C or A interacting with T are sequence driven interactions. Typically sequence driven interactions occur on the Watson-Crick face or Hoogsteen face of the nucleotide. The hybridization of two nucleic acids is affected by a number of conditions and parameters known to those of skill in the art. For example, the salt concentrations, pH, and temperature of the reaction all affect whether two nucleic acid molecules will hybridize.

Parameters for selective hybridization between two nucleic acid molecules are well known to those of skill in the art. For example, in some embodiments selective hybridization conditions can be defined as stringent hybridization conditions. For example, stringency of hybridization is controlled by both temperature and salt concentration of either or both of the hybridization and washing steps. For example, the conditions of hybridization to achieve selective hybridization can involve hybridization in high ionic strength solution (6X SSC or 6X SSPE) at a temperature that is about 12-25°C below the T_m (the melting temperature at which half of the molecules dissociate from their hybridization partners) followed by washing at a combination of temperature and salt concentration chosen so that the washing temperature is about 5°C to 20°C below the T_m. The temperature and salt conditions are readily determined empirically in preliminary experiments in which samples of reference DNA immobilized on filters are hybridized to a labeled nucleic acid of interest and then washed under conditions of different stringencies. Hybridization temperatures are typically higher for DNA-RNA and RNA-RNA hybridizations. The conditions can be used as described above to achieve stringency, or as is known in the art (Sambrook et al., *Molecular Cloning: A Laboratory Manual*, 2nd Ed., Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 1989; Kunkel et al. *Methods Enzymol.* 1987:154:367, 1987 which is herein incorporated by reference for material at least related to hybridization of nucleic acids). A preferable stringent hybridization condition for a DNA:DNA hybridization can be at about 68°C (in aqueous solution) in 6X SSC or 6X SSPE followed by washing at 68°C. Stringency of hybridization and washing, if desired, can be reduced accordingly as the degree of complementarity desired is decreased, and further, depending upon the G-C or A-T richness of any area wherein variability is searched for. Likewise, stringency of hybridization and washing, if desired, can be increased accordingly as homology desired is increased, and further, depending upon the G-C or A-T richness of any area wherein high homology is desired, all as known in the art.

Another way to define selective hybridization is by looking at the amount (percentage) of one of the nucleic acids bound to the other nucleic acid. For example, in some embodiments selective hybridization conditions would be when at least about, 60, 65, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100 percent of the limiting nucleic acid is bound to the non-limiting nucleic acid. Typically, the non-limiting nucleic acid is in for example, 10 or 100 or 1000 fold excess. This type of assay can be performed at under conditions where both the limiting and non-limiting nucleic acids are for example, 10 fold or 100 fold or 1000 fold below their k_d , or where only one of the nucleic acid molecules is 10 fold or 100 fold or 1000 fold or where one or both nucleic acid molecules are above their k_d .

Another way to define selective hybridization is by looking at the percentage of nucleic acid that gets enzymatically manipulated under conditions where hybridization is required to promote the desired enzymatic manipulation. For example, in some embodiments selective hybridization conditions would be when at least about, 60, 65, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100 percent of the nucleic acid is enzymatically manipulated under conditions which promote the enzymatic manipulation, for example if the enzymatic manipulation is DNA extension, then selective hybridization conditions would be when at least about 60, 65, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100 percent of the nucleic acid molecules are extended. Preferred conditions also include those suggested by the manufacturer or indicated in the art as being appropriate for the enzyme performing the manipulation.

Just as with homology, it is understood that there are a variety of methods herein disclosed for determining the level of hybridization between two nucleic acid molecules. It is understood that these methods and conditions can provide different percentages of hybridization between two nucleic acid molecules, but unless otherwise indicated meeting the parameters of any of the methods would be sufficient. For example if 80% hybridization was required and as long as hybridization occurs within the required parameters in any one of these methods it is considered disclosed herein.

It is understood that those of skill in the art understand that if a composition or method meets any one of these criteria for determining hybridization either collectively or singly it is a composition or method that is disclosed herein.

K. Nucleic Acids

There are a variety of molecules disclosed herein that are nucleic acid based, including, for example, riboswitches, aptamers, and nucleic acids that encode riboswitches and aptamers. The disclosed nucleic acids can be made up of for example, nucleotides, nucleotide analogs, or nucleotide substitutes. Non-limiting examples of these and other molecules are discussed herein. It is understood that for example, when a vector is expressed in a cell, that the expressed mRNA will typically be made up of A, C, G, and U. Likewise, it is understood that if a nucleic acid molecule is introduced into a cell or cell environment through for example exogenous delivery, it is advantageous that the nucleic acid molecule be made up of nucleotide analogs that reduce the degradation of the nucleic acid molecule in the cellular environment.

So long as their relevant function is maintained, riboswitches, aptamers, expression platforms and any other oligonucleotides and nucleic acids can be made up of or include modified nucleotides (nucleotide analogs). Many modified nucleotides are known and can be used in oligonucleotides and nucleic acids. A nucleotide analog is a nucleotide which contains some type of modification to either the base, sugar, or phosphate moieties. Modifications to the base moiety would include natural and synthetic modifications of A, C, G, and T/U as well as different purine or pyrimidine bases, such as uracil-5-yl, hypoxanthin-9-yl (I), and 2-aminoadenin-9-yl. A modified base includes but is not limited to 5-methylcytosine (5-me-C), 5-hydroxymethyl cytosine, xanthine, hypoxanthine, 2-aminoadenine, 6-methyl and other alkyl derivatives of adenine and guanine, 2-propyl and other alkyl derivatives of adenine and guanine, 2-thiouracil, 2-thiothymine and 2-thiocytosine, 5-halouracil and cytosine, 5-propynyl uracil and cytosine, 6-azo uracil, cytosine and thymine, 5-uracil (pseudouracil), 4-thiouracil, 8-halo, 8-amino, 8-thiol, 8-thioalkyl, 8-hydroxyl and other 8-substituted adenines and guanines, 5-halo particularly 5-bromo, 5-trifluoromethyl and other 5-substituted uracils and cytosines, 7-methylguanine and 7-methyladenine, 8-azaguanine and 8-azaadenine, 7-deazaguanine and 7-deazaadenine and 3-deazaguanine and 3-deazaadenine. Additional base modifications can be found for example in U.S. Pat. No. 3,687,808, Englisch et al., *Angewandte Chemie, International Edition*, 1991, 30, 613, and Sanghvi, Y. S., Chapter 15, *Antisense Research and Applications*, pages 289-302, Crooke, S. T. and Lebleu, B. ed., CRC Press, 1993. Certain nucleotide analogs, such as 5-substituted pyrimidines, 6-azapyrimidines and N-2, N-6 and O-6 substituted

purines, including 2-aminopropyladenine, 5-propynyluracil and 5-propynylcytosine. 5-methylcytosine can increase the stability of duplex formation. Other modified bases are those that function as universal bases. Universal bases include 3-nitropyrrole and 5-nitroindole. Universal bases substitute for the normal bases but have no bias in base pairing. That is, universal bases can base pair with any other base. Base modifications often can be combined with for example a sugar modification, such as 2'-O-methoxyethyl, to achieve unique properties such as increased duplex stability. There are numerous United States patents such as 4,845,205; 5,130,302; 5,134,066; 5,175,273; 5,367,066; 5,432,272; 5,457,187; 5,459,255; 5,484,908; 5,502,177; 5,525,711; 5,552,540; 5,587,469; 5,594,121, 5,596,091; 5,614,617; and 5,681,941, which detail and describe a range of base modifications. Each of these patents is herein incorporated by reference in its entirety, and specifically for their description of base modifications, their synthesis, their use, and their incorporation into oligonucleotides and nucleic acids.

Nucleotide analogs can also include modifications of the sugar moiety. Modifications to the sugar moiety would include natural modifications of the ribose and deoxyribose as well as synthetic modifications. Sugar modifications include but are not limited to the following modifications at the 2' position: OH; F; O-, S-, or N-alkyl; O-, S-, or N-alkenyl; O-, S- or N-alkynyl; or O-alkyl-O-alkyl, wherein the alkyl, alkenyl and alkynyl can be substituted or unsubstituted C1 to C10, alkyl or C2 to C10 alkenyl and alkynyl. 2' sugar modifications also include but are not limited to -O[(CH₂)_nO]_mCH₃, -O(CH₂)_nOCH₃, -O(CH₂)_nNH₂, -O(CH₂)_nCH₃, -O(CH₂)_n-ONH₂, and -O(CH₂)_nON[(CH₂)_nCH₃]₂, where n and m are from 1 to about 10.

Other modifications at the 2' position include but are not limited to: C1 to C10 lower alkyl, substituted lower alkyl, alkaryl, aralkyl, O-alkaryl or O-aralkyl, SH, SCH₃, OCN, Cl, Br, CN, CF₃, OCF₃, SOCH₃, SO₂CH₃, ONO₂, NO₂, N₃, NH₂, heterocycloalkyl, heterocycloalkaryl, aminoalkylamino, polyalkylamino, substituted silyl, an RNA cleaving group, a reporter group, an intercalator, a group for improving the pharmacokinetic properties of an oligonucleotide, or a group for improving the pharmacodynamic properties of an oligonucleotide, and other substituents having similar properties. Similar modifications can also be made at other positions on the sugar, particularly the 3' position of the sugar on the 3' terminal nucleotide or in 2'-5' linked oligonucleotides and the 5' position of 5' terminal nucleotide. Modified sugars would also include those that contain modifications at the bridging ring oxygen, such as CH₂

and S. Nucleotide sugar analogs can also have sugar mimetics such as cyclobutyl moieties in place of the pentofuranosyl sugar. There are numerous United States patents that teach the preparation of such modified sugar structures such as 4,981,957; 5,118,800; 5,319,080; 5,359,044; 5,393,878; 5,446,137; 5,466,786; 5,514,785; 5,519,134; 5,567,811; 5,576,427; 5,591,722; 5,597,909; 5,610,300; 5,627,053; 5,639,873; 5,646,265; 5,658,873; 5,670,633; and 5,700,920, each of which is herein incorporated by reference in its entirety, and specifically for their description of modified sugar structures, their synthesis, their use, and their incorporation into nucleotides, oligonucleotides and nucleic acids.

Nucleotide analogs can also be modified at the phosphate moiety. Modified phosphate moieties include but are not limited to those that can be modified so that the linkage between two nucleotides contains a phosphorothioate, chiral phosphorothioate, phosphorodithioate, phosphotriester, aminoalkylphosphotriester, methyl and other alkyl phosphonates including 3'-alkylene phosphonate and chiral phosphonates, phosphinates, phosphoramidates including 3'-amino phosphoramidate and aminoalkylphosphoramidates, thionophosphoramidates, thionoalkylphosphonates, thionoalkylphosphotriesters, and boranophosphates. It is understood that these phosphate or modified phosphate linkages between two nucleotides can be through a 3'-5' linkage or a 2'-5' linkage, and the linkage can contain inverted polarity such as 3'-5' to 5'-3' or 2'-5' to 5'-2'. Various salts, mixed salts and free acid forms are also included. Numerous United States patents teach how to make and use nucleotides containing modified phosphates and include but are not limited to, 3,687,808; 4,469,863; 4,476,301; 5,023,243; 5,177,196; 5,188,897; 5,264,423; 5,276,019; 5,278,302; 5,286,717; 5,321,131; 5,399,676; 5,405,939; 5,453,496; 5,455,233; 5,466,677; 5,476,925; 5,519,126; 5,536,821; 5,541,306; 5,550,111; 5,563,253; 5,571,799; 5,587,361; and 5,625,050, each of which is herein incorporated by reference its entirety, and specifically for their description of modified phosphates, their synthesis, their use, and their incorporation into nucleotides, oligonucleotides and nucleic acids.

It is understood that nucleotide analogs need only contain a single modification, but can also contain multiple modifications within one of the moieties or between different moieties.

Nucleotide substitutes are molecules having similar functional properties to nucleotides, but which do not contain a phosphate moiety, such as peptide nucleic acid

(PNA). Nucleotide substitutes are molecules that will recognize and hybridize to (base pair to) complementary nucleic acids in a Watson-Crick or Hoogsteen manner, but which are linked together through a moiety other than a phosphate moiety. Nucleotide substitutes are able to conform to a double helix type structure when interacting with the appropriate target nucleic acid.

Nucleotide substitutes are nucleotides or nucleotide analogs that have had the phosphate moiety and/or sugar moieties replaced. Nucleotide substitutes do not contain a standard phosphorus atom. Substitutes for the phosphate can be for example, short chain alkyl or cycloalkyl internucleoside linkages, mixed heteroatom and alkyl or cycloalkyl internucleoside linkages, or one or more short chain heteroatomic or heterocyclic internucleoside linkages. These include those having morpholino linkages (formed in part from the sugar portion of a nucleoside); siloxane backbones; sulfide, sulfoxide and sulfone backbones; formacetyl and thioformacetyl backbones; methylene formacetyl and thioformacetyl backbones; alkene containing backbones; sulfamate backbones; methyleneimino and methylenehydrazino backbones; sulfonate and sulfonamide backbones; amide backbones; and others having mixed N, O, S and CH₂ component parts. Numerous United States patents disclose how to make and use these types of phosphate replacements and include but are not limited to 5,034,506; 5,166,315; 5,185,444; 5,214,134; 5,216,141; 5,235,033; 5,264,562; 5,264,564; 5,405,938; 5,434,257; 5,466,677; 5,470,967; 5,489,677; 5,541,307; 5,561,225; 5,596,086; 5,602,240; 5,610,289; 5,602,240; 5,608,046; 5,610,289; 5,618,704; 5,623,070; 5,663,312; 5,633,360; 5,677,437; and 5,677,439, each of which is herein incorporated by reference its entirety, and specifically for their description of phosphate replacements, their synthesis, their use, and their incorporation into nucleotides, oligonucleotides and nucleic acids.

It is also understood in a nucleotide substitute that both the sugar and the phosphate moieties of the nucleotide can be replaced, by for example an amide type linkage (aminoethylglycine) (PNA). United States patents 5,539,082; 5,714,331; and 5,719,262 teach how to make and use PNA molecules, each of which is herein incorporated by reference. (See also Nielsen *et al.*, *Science* 254:1497-1500 (1991)).

Oligonucleotides and nucleic acids can be comprised of nucleotides and can be made up of different types of nucleotides or the same type of nucleotides. For example, one or more of the nucleotides in an oligonucleotide can be ribonucleotides, 2'-O-methyl

ribonucleotides, or a mixture of ribonucleotides and 2'-O-methyl ribonucleotides; about 10% to about 50% of the nucleotides can be ribonucleotides, 2'-O-methyl ribonucleotides, or a mixture of ribonucleotides and 2'-O-methyl ribonucleotides; about 50% or more of the nucleotides can be ribonucleotides, 2'-O-methyl ribonucleotides, or a mixture of ribonucleotides and 2'-O-methyl ribonucleotides; or all of the nucleotides are ribonucleotides, 2'-O-methyl ribonucleotides, or a mixture of ribonucleotides and 2'-O-methyl ribonucleotides. Such oligonucleotides and nucleic acids can be referred to as chimeric oligonucleotides and chimeric nucleic acids.

L. Solid Supports

10 Solid supports are solid-state substrates or supports with which molecules (such as trigger molecules) and riboswitches (or other components used in, or produced by, the disclosed methods) can be associated. Riboswitches and other molecules can be associated with solid supports directly or indirectly. For example, analytes (e.g., trigger molecules, test compounds) can be bound to the surface of a solid support or associated with capture agents (e.g., compounds or molecules that bind an analyte) immobilized on solid supports. As another example, riboswitches can be bound to the surface of a solid support or associated with probes immobilized on solid supports. An array is a solid support to which multiple riboswitches, probes or other molecules have been associated in an array, grid, or other organized pattern.

20 Solid-state substrates for use in solid supports can include any solid material with which components can be associated, directly or indirectly. This includes materials such as acrylamide, agarose, cellulose, nitrocellulose, glass, gold, polystyrene, polyethylene vinyl acetate, polypropylene, polymethacrylate, polyethylene, polyethylene oxide, polysilicates, polycarbonates, teflon, fluorocarbons, nylon, silicon rubber, polyanhydrides, polyglycolic acid, polylactic acid, polyorthoesters, functionalized silane, polypropylfumerate, collagen, glycosaminoglycans, and polyamino acids. Solid-state substrates can have any useful form including thin film, membrane, bottles, dishes, fibers, woven fibers, shaped polymers, particles, beads, microparticles, or a combination. Solid-state substrates and solid supports can be porous or non-porous. A chip is a rectangular or square small piece of material. Preferred forms for solid-state substrates are thin films, beads, or chips. A useful form for a solid-state substrate is a microtiter dish. In some embodiments, a multiwell glass slide can be employed.

An array can include a plurality of riboswitches, trigger molecules, other molecules, compounds or probes immobilized at identified or predefined locations on the solid support. Each predefined location on the solid support generally has one type of component (that is, all the components at that location are the same). Alternatively,
5 multiple types of components can be immobilized in the same predefined location on a solid support. Each location will have multiple copies of the given components. The spatial separation of different components on the solid support allows separate detection and identification.

Although useful, it is not required that the solid support be a single unit or
10 structure. A set of riboswitches, trigger molecules, other molecules, compounds and/or probes can be distributed over any number of solid supports. For example, at one extreme, each component can be immobilized in a separate reaction tube or container, or on separate beads or microparticles.

Methods for immobilization of oligonucleotides to solid-state substrates are well
15 established. Oligonucleotides, including address probes and detection probes, can be coupled to substrates using established coupling methods. For example, suitable attachment methods are described by Pease *et al.*, *Proc. Natl. Acad. Sci. USA* 91(11):5022-5026 (1994), and Khrapko *et al.*, *Mol Biol (Mosk) (USSR)* 25:718-730 (1991). A method for immobilization of 3'-amine oligonucleotides on casein-coated
20 slides is described by Stimpson *et al.*, *Proc. Natl. Acad. Sci. USA* 92:6379-6383 (1995). A useful method of attaching oligonucleotides to solid-state substrates is described by Guo *et al.*, *Nucleic Acids Res.* 22:5456-5465 (1994).

Each of the components (for example, riboswitches, trigger molecules, or other molecules) immobilized on the solid support can be located in a different predefined
25 region of the solid support. The different locations can be different reaction chambers. Each of the different predefined regions can be physically separated from each other of the different regions. The distance between the different predefined regions of the solid support can be either fixed or variable. For example, in an array, each of the components can be arranged at fixed distances from each other, while components associated with
30 beads will not be in a fixed spatial relationship. In particular, the use of multiple solid support units (for example, multiple beads) will result in variable distances.

Components can be associated or immobilized on a solid support at any density. Components can be immobilized to the solid support at a density exceeding 400 different

components per cubic centimeter. Arrays of components can have any number of components. For example, an array can have at least 1,000 different components immobilized on the solid support, at least 10,000 different components immobilized on the solid support, at least 100,000 different components immobilized on the solid support, or at least 1,000,000 different components immobilized on the solid support.

M. Kits

The materials described above as well as other materials can be packaged together in any suitable combination as a kit useful for performing, or aiding in the performance of, the disclosed method. It is useful if the kit components in a given kit are designed and adapted for use together in the disclosed method. For example disclosed are kits for detecting compounds, the kit comprising one or more biosensor riboswitches. The kits also can contain reagents and labels for detecting activation of the riboswitches.

N. Mixtures

Disclosed are mixtures formed by performing or preparing to perform the disclosed method. For example, disclosed are mixtures comprising riboswitches and trigger molecules.

Whenever the method involves mixing or bringing into contact compositions or components or reagents, performing the method creates a number of different mixtures. For example, if the method includes 3 mixing steps, after each one of these steps a unique mixture is formed if the steps are performed separately. In addition, a mixture is formed at the completion of all of the steps regardless of how the steps were performed. The present disclosure contemplates these mixtures, obtained by the performance of the disclosed methods as well as mixtures containing any disclosed reagent, composition, or component, for example, disclosed herein.

O. Systems

Disclosed are systems useful for performing, or aiding in the performance of, the disclosed method. Systems generally comprise combinations of articles of manufacture such as structures, machines, devices, and the like, and compositions, compounds, materials, and the like. Such combinations that are disclosed or that are apparent from the disclosure are contemplated. For example, disclosed and contemplated are systems comprising iosensor riboswitches, a solid support and a signal-reading device.

P. Data Structures and Computer Control

Disclosed are data structures used in, generated by, or generated from, the disclosed method. Data structures generally are any form of data, information, and/or objects collected, organized, stored, and/or embodied in a composition or medium.

5 Riboswitch structures and activation measurements stored in electronic form, such as in RAM or on a storage disk, is a type of data structure.

The disclosed method, or any part thereof or preparation therefor, can be controlled, managed, or otherwise assisted by computer control. Such computer control can be accomplished by a computer controlled process or method, can use and/or
10 generate data structures, and can use a computer program. Such computer control, computer controlled processes, data structures, and computer programs are contemplated and should be understood to be disclosed herein.

Methods

Disclosed are methods for activating, deactivating or blocking a riboswitch. Such
15 methods can involve, for example, bringing into contact a riboswitch and a compound or trigger molecule that can activate, deactivate or block the riboswitch. Riboswitches function to control gene expression through the binding or removal of a trigger molecule. Compounds can be used to activate, deactivate or block a riboswitch. The trigger molecule for a riboswitch (as well as other activating compounds) can be used to activate
20 a riboswitch. Compounds other than the trigger molecule generally can be used to deactivate or block a riboswitch. Riboswitches can also be deactivated by, for example, removing trigger molecules from the presence of the riboswitch. Thus, the disclosed method of deactivating a riboswitch can involve, for example, removing a trigger molecule (or other activating compound) from the presence or contact with the
25 riboswitch. A riboswitch can be blocked by, for example, binding of an analog of the trigger molecule that does not activate the riboswitch.

Also disclosed are methods for altering expression of an RNA molecule, or of a gene encoding an RNA molecule, where the RNA molecule includes a riboswitch, by bringing a compound into contact with the RNA molecule. Riboswitches function to
30 control gene expression through the binding or removal of a trigger molecule. Thus, subjecting an RNA molecule of interest that includes a riboswitch to conditions that activate, deactivate or block the riboswitch can be used to alter expression of the RNA. Expression can be altered as a result of, for example, termination of transcription or

blocking of ribosome binding to the RNA. Binding of a trigger molecule can, depending on the nature of the riboswitch, reduce or prevent expression of the RNA molecule or promote or increase expression of the RNA molecule.

Also disclosed are methods for regulating expression of an RNA molecule, or of
5 a gene encoding an RNA molecule, by operably linking a riboswitch to the RNA molecule. A riboswitch can be operably linked to an RNA molecule in any suitable manner, including, for example, by physically joining the riboswitch to the RNA molecule or by engineering nucleic acid encoding the RNA molecule to include and encode the riboswitch such that the RNA produced from the engineered nucleic acid has
10 the riboswitch operably linked to the RNA molecule. Subjecting a riboswitch operably linked to an RNA molecule of interest to conditions that activate, deactivate or block the riboswitch can be used to alter expression of the RNA.

Also disclosed are methods for regulating expression of a naturally occurring gene or RNA that contains a riboswitch by activating, deactivating or blocking the
15 riboswitch. If the gene is essential for survival of a cell or organism that harbors it, activating, deactivating or blocking the riboswitch can in death, stasis or debilitation of the cell or organism. For example, activating a naturally occurring riboswitch in a naturally occurring gene that is essential to survival of a microorganism can result in death of the microorganism (if activation of the riboswitch turns off or represses
20 expression). This is one basis for the use of the disclosed compounds and methods for antimicrobial and antibiotic effects.

Also disclosed are methods for regulating expression of an isolated, engineered or recombinant gene or RNA that contains a riboswitch by activating, deactivating or blocking the riboswitch. The gene or RNA can be engineered or can be recombinant in
25 any manner. For example, the riboswitch and coding region of the RNA can be heterologous, the riboswitch can be recombinant or chimeric, or both. If the gene encodes a desired expression product, activating or deactivating the riboswitch can be used to induce expression of the gene and thus result in production of the expression product. If the gene encodes an inducer or repressor of gene expression or of another
30 cellular process, activation, deactivation or blocking of the riboswitch can result in induction, repression, or de-repression of other, regulated genes or cellular processes. Many such secondary regulatory effects are known and can be adapted for use with

riboswitches. An advantage of riboswitches as the primary control for such regulation is that riboswitch trigger molecules can be small, non-antigenic molecules.

Also disclosed are methods for altering the regulation of a riboswitch by operably linking an aptamer domain to the expression platform domain of the riboswitch (which is a chimeric riboswitch). The aptamer domain can then mediate regulation of the riboswitch through the action of, for example, a trigger molecule for the aptamer domain. Aptamer domains can be operably linked to expression platform domains of riboswitches in any suitable manner, including, for example, by replacing the normal or natural aptamer domain of the riboswitch with the new aptamer domain. Generally, any compound or condition that can activate, deactivate or block the riboswitch from which the aptamer domain is derived can be used to activate, deactivate or block the chimeric riboswitch.

Also disclosed are methods for inactivating a riboswitch by covalently altering the riboswitch (by, for example, crosslinking parts of the riboswitch or coupling a compound to the riboswitch). Inactivation of a riboswitch in this manner can result from, for example, an alteration that prevents the trigger molecule for the riboswitch from binding, that prevents the change in state of the riboswitch upon binding of the trigger molecule, or that prevents the expression platform domain of the riboswitch from affecting expression upon binding of the trigger molecule.

Also disclosed are methods for selecting, designing or deriving new riboswitches and/or new aptamers that recognize new trigger molecules. Such methods can involve production of a set of aptamer variants in a riboswitch, assessing the activation of the variant riboswitches in the presence of a compound of interest, selecting variant riboswitches that were activated (or, for example, the riboswitches that were the most highly or the most selectively activated), and repeating these steps until a variant riboswitch of a desired activity, specificity, combination of activity and specificity, or other combination of properties results. Also disclosed are riboswitches and aptamer domains produced by these methods.

Techniques for in vitro selection and in vitro evolution of functional nucleic acid molecules are known and can be adapted for use with riboswitches and their components. Useful techniques are described by, for example, A. Roth and R. R. Breaker (2003) Selection in vitro of allosteric ribozymes. In: Methods in Molecular Biology Series - Catalytic Nucleic Acid Protocols (Sioud, M., ed.), Humana, Totowa,

- NJ; R. R. Breaker (2002) Engineered Allosteric Ribozymes as Biosensor Components. *Curr. Opin. Biotechnol.* 13:31-39; G. M. Emilsson and R. R. Breaker (2002) Deoxyribozymes: New Activities and New Applications. *Cell. Mol. Life Sci.* 59:596-607; Y. Li, R. R. Breaker (2001) In vitro Selection of Kinase and Ligase
- 5 Deoxyribozymes. *Methods* 23:179-190; G. A. Soukup, R. R. Breaker (2000) Allosteric Ribozymes. In: *Ribozymes: Biology and Biotechnology*. R. K. Gaur and G. Krupp eds. Eaton Publishing; G. A. Soukup, R. R. Breaker (2000) Allosteric Nucleic Acid Catalysts. *Curr. Opin. Struct. Biol.* 10:318-325; G. A. Soukup, R. R. Breaker (1999) Nucleic Acid Molecular Switches. *Trends Biotechnol.* 17:469-476; R. R. Breaker (1999) In vitro
- 10 Selection of Self-cleaving Ribozymes and Deoxyribozymes. In: *Intracellular Ribozyme Applications: Principles and Protocols*. L. Couture, J. Rossi eds. Horizon Scientific Press, Norfolk, England; R. R. Breaker (1997) In vitro Selection of Catalytic Polynucleotides. *Chem. Rev.* 97:371-390; and references cited therein; each of these publications being specifically incorporated herein by reference for their description of in
- 15 vitro selections and evolution techniques.

Also disclosed are methods for selecting and identifying compounds that can activate, deactivate or block a riboswitch. Activation of a riboswitch refers to the change in state of the riboswitch upon binding of a trigger molecule. A riboswitch can be activated by compounds other than the trigger molecule and in ways other than binding

20 of a trigger molecule. The term trigger molecule is used herein to refer to molecules and compounds that can activate a riboswitch. This includes the natural or normal trigger molecule for the riboswitch and other compounds that can activate the riboswitch. Natural or normal trigger molecules are the trigger molecule for a given riboswitch in nature or, in the case of some non-natural riboswitches, the trigger molecule for which

25 the riboswitch was designed or with which the riboswitch was selected (as in, for example, in vitro selection or in vitro evolution techniques). Non-natural trigger molecules can be referred to as non-natural trigger molecules.

Deactivation of a riboswitch refers to the change in state of the riboswitch when the trigger molecule is not bound. A riboswitch can be deactivated by binding of

30 compounds other than the trigger molecule and in ways other than removal of the trigger molecule. Blocking of a riboswitch refers to a condition or state of the riboswitch where the presence of the trigger molecule does not activate the riboswitch.

Also disclosed are methods of identifying compounds that activate, deactivate or block a riboswitch. For examples, compounds that activate a riboswitch can be identified by bringing into contact a test compound and a riboswitch and assessing activation of the riboswitch. If the riboswitch is activated, the test compound is
5 identified as a compound that activates the riboswitch. Activation of a riboswitch can be assessed in any suitable manner. For example, the riboswitch can be linked to a reporter RNA and expression, expression level, or change in expression level of the reporter RNA can be measured in the presence and absence of the test compound. As another example, the riboswitch can include a conformation dependent label, the signal from which
10 changes depending on the activation state of the riboswitch. Such a riboswitch preferably uses an aptamer domain from or derived from a naturally occurring riboswitch. As can be seen, assessment of activation of a riboswitch can be performed with the use of a control assay or measurement or without the use of a control assay or measurement. Methods for identifying compounds that deactivate a riboswitch can be
15 performed in analogous ways.

Identification of compounds that block a riboswitch can be accomplished in any suitable manner. For example, an assay can be performed for assessing activation or deactivation of a riboswitch in the presence of a compound known to activate or deactivate the riboswitch and in the presence of a test compound. If activation or
20 deactivation is not observed as would be observed in the absence of the test compound, then the test compound is identified as a compound that blocks activation or deactivation of the riboswitch.

Also disclosed are methods of detecting compounds using biosensor riboswitches. The method can include bringing into contact a test sample and a
25 biosensor riboswitch and assessing the activation of the biosensor riboswitch. Activation of the biosensor riboswitch indicates the presence of the trigger molecule for the biosensor riboswitch in the test sample. Biosensor riboswitches are engineered riboswitches that produce a detectable signal in the presence of their cognate trigger molecule. Useful biosensor riboswitches can be triggered at or above threshold levels of
30 the trigger molecules. Biosensor riboswitches can be designed for use in vivo or in vitro. For example, biosensor riboswitches operably linked to a reporter RNA that encodes a protein that serves as or is involved in producing a signal can be used in vivo by engineering a cell or organism to harbor a nucleic acid construct encoding the

riboswitch/reporter RNA. An example of a biosensor riboswitch for use in vitro is a riboswitch that includes a conformation dependent label, the signal from which changes depending on the activation state of the riboswitch. Such a biosensor riboswitch preferably uses an aptamer domain from or derived from a naturally occurring
5 riboswitch.

Biosensor riboswitches can be used to monitor changing conditions because riboswitch activation is reversible when the concentration of the trigger molecule falls and so the signal can vary as concentration of the trigger molecule varies. The range of concentration of trigger molecules that can be detected can be varied by engineering
10 riboswitches having different dissociation constants for the trigger molecule. This can easily be accomplished by, for example, "degrading" the sensitivity of a riboswitch having high affinity for the trigger molecule. A range of concentrations can be monitored by using multiple biosensor riboswitches of different sensitivities in the same sensor or assay.

Also disclosed are compounds made by identifying a compound that activates, deactivates or blocks a riboswitch and manufacturing the identified compound. This can be accomplished by, for example, combining compound identification methods as disclosed elsewhere herein with methods for manufacturing the identified compounds. For example, compounds can be made by bringing into contact a test compound and a
20 riboswitch, assessing activation of the riboswitch, and, if the riboswitch is activated by the test compound, manufacturing the test compound that activates the riboswitch as the compound.

Also disclosed are compounds made by checking activation, deactivation or blocking of a riboswitch by a compound and manufacturing the checked compound.
25 This can be accomplished by, for example, combining compound activation, deactivation or blocking assessment methods as disclosed elsewhere herein with methods for manufacturing the checked compounds. For example, compounds can be made by bringing into contact a test compound and a riboswitch, assessing activation of the riboswitch, and, if the riboswitch is activated by the test compound, manufacturing the
30 test compound that activates the riboswitch as the compound. Checking compounds for their ability to activate, deactivate or block a riboswitch refers to both identification of compounds previously unknown to activate, deactivate or block a riboswitch and to

assessing the ability of a compound to activate, deactivate or block a riboswitch where the compound was already known to activate, deactivate or block the riboswitch.

Disclosed is a method of detecting a compound of interest, the method comprising bringing into contact a sample and a riboswitch, wherein the riboswitch is activated by the compound of interest, wherein the riboswitch produces a signal when activated by the compound of interest, wherein the riboswitch produces a signal when the sample contains the compound of interest. The riboswitch can change conformation when activated by the compound of interest, wherein the change in conformation produces a signal via a conformation dependent label. The riboswitch can change conformation when activated by the compound of interest, wherein the change in conformation causes a change in expression of an RNA linked to the riboswitch, wherein the change in expression produces a signal. The signal can be produced by a reporter protein expressed from the RNA linked to the riboswitch.

Disclosed is a method comprising (a) testing a compound for inhibition of gene expression of a gene encoding an RNA comprising a riboswitch, wherein the inhibition is via the riboswitch, and (b) inhibiting gene expression by bringing into contact a cell and a compound that inhibited gene expression in step (a), wherein the cell comprises a gene encoding an RNA comprising a riboswitch, wherein the compound inhibits expression of the gene by binding to the riboswitch.

Also disclosed is a method of identifying riboswitches, the method comprising assessing in-line spontaneous cleavage of an RNA molecule in the presence and absence of a compound, wherein the RNA molecule is encoded by a gene regulated by the compound, wherein a change in the pattern of in-line spontaneous cleavage of the RNA molecule indicates a riboswitch.

A. Identification of Antimicrobial Compounds

Riboswitches are a new class of structured RNAs that have evolved for the purpose of binding small organic molecules. The natural binding pocket of riboswitches can be targeted with metabolite analogs or by compounds that mimic the shape-space of the natural metabolite. Riboswitches are: (1) found in numerous Gram-positive and Gram-negative bacteria including *Bacillus anthracis*, (2) fundamental regulators of gene expression in these bacteria, (3) present in multiple copies that would be unlikely to evolve simultaneous resistance, and (4) not yet proven to exist in humans. This combination of features make riboswitches attractive targets for new antimicrobial

compounds. Further, the small molecule ligands of riboswitches provide useful sites for derivitization to produce drug candidates.

Once a class of riboswitch has been identified and its potential as a drug target assessed (by, for example, determining how many genes in a target organism are regulated by that class of riboswitch), candidate molecules can be identified. The following provides an illustration of this using the SAM riboswitch (see Example 7).

SAM analogs that substitute the reactive methyl and sulfonium ion center with stable sulfur-based linkages (YBD-2 and YBD3) are recognized with adequate affinity (low to mid-nanomolar range) by the riboswitch to serve as a platform for synthesis of additional SAM analogs. In addition, a wider range of linkage analogs (N- and C-based linkages) can be synthesized and tested to provide the optimal platform upon which to make amino acid and nucleoside derivations.

Sulfoxide and sulfone derivatives of SAM can be used to generate analogs. Established synthetic protocols described in Ronald T. Borchardt and Yih Shiong Wu, Potential inhibitor of S-adenosylmethionine-dependent methyltransferase. 1. Modification of the amino acid portion of S-adenosylhomocysteine. J. Med. Chem. 17, 862-868, 1974, can be used, for example. These and other analogs can be synthesized and assayed for binding sequentially or in small groups. Additional SAM analogs can be designed during the progression of compound identification based on the recognition determinants that are established in each round. Simple binding assays can be conducted on *B. subtilis* and *B. anthracis* riboswitch RNAs as described elsewhere herein. More advanced assays can also be used.

The most promising SAM analog lead compounds must enter bacterial cells and bind riboswitches while remaining metabolically inert. In addition, useful SAM analogs must be bound tightly by the riboswitch, but must also fail to compete for SAM in the active sites of protein enzymes, or there is a risk of generating an undesirable toxic effect in the patient's cells. As a preliminary assessment of these issues, compounds can be tested for their ability to disrupt *B. subtilis* growth, but fail to affect *E. coli* cultures (which use SAM but lack SAM riboswitches). To screen for lead compound candidates, parallel bacterial cultures can be grown as follows:

1. *B. subtilis* can be cultured in glucose minimal media in the absence of exogenously supplied SAM analogs.

2. *B. subtilis* can be cultured in glucose minimal media in the presence of exogenously supplied SAM analogs (high doses can be selected, to be followed by repeated experiments designed to test a concentration range of the putative drug compound).

5 3. *E. coli* can be cultured in glucose minimal media in the presence of exogenously supplied SAM analogs (high doses will be selected, to be followed by repeated experiments designed to test a concentration range of the putative drug compound).

10 Fitness of the various cultures can be compared by measurement of cellular doubling times. A range of concentrations for the drug compounds can be tested using cultures grown in microtiter plates and analyzed using a microplate reader from another laboratory. Culture 1 is expected to grow well. Drugs that inhibit culture 2 may or may not inhibit growth of culture 3. Drugs that similarly inhibit both culture 2 and culture 3 upon exposure to a wide range of drug concentrations can reflect general toxicity
15 induced by the exogenous compound (*i.e.*, inhibition of many different cellular processes, in addition or in place of riboswitch inhibition). Successful drug candidates identified in this screen will inhibit *E. coli* only at very high doses, if at all, and will inhibit *B. subtilis* at much (>10-fold) lower concentrations.

20 As derivization points on SAM are identified, efficient identification of lead drug compounds will require larger-scale screening of appropriate SAM analogs or generic chemical libraries. A high-throughput screen can be created by one or two different methods using nucleic acid engineering principles. Adaptation of both fluorescent sensor designs outlined below to formats that are compatible with high-throughput screening assays can be accommodated by using immobilization methods or solution-based
25 methods.

One way to create a reporter is to add a third function to the riboswitch by adding a domain that catalyzes the release of a fluorescent tag upon SAM binding to the riboswitch domain. In the final reporter construct, this catalytic domain can be linked to the *yitJ* SAM riboswitch through a communication module that relays the ligand binding
30 event by allowing the correct folding of the catalytic domain for generating the fluorescent signal. This can be accomplished as outlined below.

SAM RiboReporter Pool Design: A DNA template for in vitro transcription to RNA (Fig. 10) has been constructed by PCR amplification using the appropriate DNA

template and primer sequences. In this construct, stem II of the hammerhead (stem P1 of the SAM aptamer) has been randomized to present more than 250 million possible sequence combinations, wherein some inevitably will permit function of the ribozyme only when the aptamer is occupied by SAM or a related high-affinity analog. Each molecule in the population of constructs is identical in sequence except at the random domain where multiple copies of every possible combination of sequence will be represented in the population.

SAM RiboReporter Selection: The *in vitro* selection protocol can be a repetitive iteration of the following steps:

1. Transcribe RNA *in vitro* by standard methods. Include [α - 32 P] UTP to incorporate radioactivity throughout the RNA.
2. Purify full length RNA on denaturing PAGE by standard methods.
3. Incubate full length RNA (~100 pmoles) in negative selection buffer containing sufficient magnesium for catalytic activity (20 mM) but no SAM. Incubate 4 h at room temperature (~23°C), with thermocycling or alkaline denaturation as needed to preclude the emergence of selfish molecules.
4. Purify full length RNA on denaturing PAGE and discard RNAs that react in the absence of SAM.
5. Incubate in positive selection buffer containing 20 mM Mg $^{2+}$ and SAM (pH 7.5 at 23°C). Incubate 20 min at room temperature.
6. Purify cleaved RNA on denaturing PAGE to recover switches that bound SAM and allowed self-cleavage of the RNA.
7. Reverse transcribe RNA to DNA.
8. PCR amplify DNA with primers that reintroduced cleaved portion of RNA.

The concentration of SAM in step 4 can be 100 μ M initially and can be reduced as the selection proceeds. The progress of recovering successful communication modules can be assessed by the amount of cleavage observed on the purification gel in step 6. The selection endpoint can be either when the population approaches 100% cleavage in 10 nM SAM (conditions for maximal activity of the parental ribozyme and riboswitch) or when the population approaches a plateau in activity that does not improve over multiple rounds. The end population can then be sequenced. Individual communication module

clones can be assayed for generation of a fluorescent signal in the screening construct in the presence of SAM.

A fluorescent signal can also be generated by riboswitch-mediated triggering of a molecular beacon. In this design, riboswitch conformational changes cause a folded molecular beacon tagged with both a fluor and a quencher to unfold and force the fluor away from the quencher by forming a helix with the riboswitch. This mechanism is easy to adapt to existing riboswitches, as this method can take advantage of the ligand-mediated formation of terminator and anti-terminator stems that are involved in transcription control.

To use riboswitches to report ligand binding by binding a molecular beacon, the appropriate construct must be determined empirically. The optimum length and nucleotide composition of the molecular beacon and its binding site on the riboswitch can be tested systematically to result in the highest signal-to-noise ratio. The validity of the assay can be determined by comparing apparent relative binding affinities of different SAM analogs to a molecular beacon-coupled riboswitch (determined by rate of fluorescent signal generation) to the binding constants determined by standard in-line probing.

Examples

A. Example 1: Coenzyme B₁₂ (AdoCbl) Riboswitches

The example described testing and analysis of a riboswitch that controls gene expression by binding coenzyme B₁₂.

1. Methods

i. Chemicals and Oligonucleotides

Coenzyme B₁₂ (5'-deoxy-5'-adenosylcobalamin or "AdoCbl") and its analogs methylcobalamin, cobinamide dicyanide, and cyanocobalamin were purchased from Sigma. Tritiated AdoCbl was prepared as described previously (Brown and Zou, Thermolysis of coenzymes B₁₂ at physiological temperatures: activation parameters for cobalt-carbon bond homolysis and a quantitative analysis of the perturbation of the homolysis equilibrium by the ribonucleoside triphosphate reductase from *Lactobacillus leichmannii*. J. Inorg. Biochem. 77, 185-195 (1999)). For information regarding the AdoCbl analogs B⁶,N⁶-dimethyl-AdoCbl, N⁶-methyl-AdoCbl, N¹-methyl-AdoCbl, 3-

deaza-AdoCbl, PurCbl, 2'-deoxy-AdoCbl and 13-epi-AdoCbl, see Toraya, In: Chemistry and Biochemistry of B₁₂. Banerjee, R. Ed. (Wiley, New York) pp. 783-809 (1999).

DNA oligonucleotides were synthesized by the Keck Foundation Biotechnology Resource Center at Yale University. DNAs were purified by denaturing (8 M urea) PAGE and isolated from the gel by crush/soaking in 10 mM Tris-HCl (pH 7.5 at 23°C), 200 mM NaCl and 1 mM EDTA. The DNA was recovered from the solution by precipitation with ethanol, resuspended in water and stored at -20°C until use.

ii. RNA Structure Analysis by In-line probing

Precursor mRNA leader molecules were prepared by *in vitro* transcription from templates generated by PCR (see In vivo Expression Constructs and Assays section below) and 5' ³²P-labeled using methods described previously (Soukup and Breaker, Allosteric nucleic acid catalysts. Curr. Opin. Struct. Biol. 10, 318-325 (2000)). Approximately 20 nM of labeled RNA precursor was incubated as described in the brief description of Figure 1. Accompanying digestions were carried out using reaction conditions similar to those described previously (Soukup and Breaker, Relationship between internucleotide linkage geometry and the stability of P_i-NA. RNA 5, 1308-1325 (1999)). To prevent light-induced degradation of ligands, incubations were protected from exposure to light by wrapping each tube with aluminum foil.

iii. Equilibrium Dialysis Assays

Each equilibrium dialysis experiment was conducted using a Dispo-Equilibrium Dialyzer (ED-1, Harvard Bioscience) apparatus, wherein two chambers (a and b) each contained 25 µL of equilibration buffer (50 mM Tris-HCl [pH 8.3 at 25°C], 20 mM MgCl₂). The chambers were separated by a dialysis membrane with a 5,000 Dalton molecular weight cut-off. In each experiment (I- IV, boxed), 100 pmoles of ³H-AdoCbl were included in chamber a, and other additives were included as designated (+) for each chamber. In each step, equilibrations were allowed to proceed for 10 hrs at 25°C before samples were quantitated or before subsequent manipulations were carried out. Quantitation was achieved by liquid scintillation counting using 5 or 10 µL of solution from each chamber.

Dialysis samples were protected from exposure to light by wrapping each apparatus with aluminum foil.

iv. *In vivo* Expression Constructs and Assays

E. coli K-12 strain was used for all *btuB-lacZ* expression assays and Top10 cells (Invitrogen) were used for plasmid preparation. A DNA (nucleotides -70 to 450) encompassing the *btuB* leader sequence was amplified as an EcoRI-BamHI fragment by colony PCR from *E. coli* strain MC4100 (a gift from S. Gottesman, NIH). The wild-type construct and mutant constructs were inserted into plasmid pRS414 (a gift from R. Simons, UCLA; Simons et al., Improved single and multicopy lac-based cloning vectors for protein and operon fusions. *Gene* 53, 85-96 (1987)), in frame with the 9th codon of *lacZ* (β -galactosidase). Mutant constructs were generated by a three-step PCR strategy wherein regions upstream and down stream of the mutation site were amplified separately with the appropriate DNA primers that introduced the desired sequence changes. The resulting fragments were purified by agarose gel electrophoresis, and then combined and amplified by PCR using primers that correspond to the ends of the full-length construct. The resulting constructs were cloned and sequenced. Constructs whose sequence was confirmed were used for expression analysis and were used as templates for subsequent preparation of PCR-derived DNAs for *in vitro* transcription.

The in-frame fusions between various *btuB* leader sequences and *lacZ* generated as described above were used to determine the levels of expression by employing a α /3-galactosidase assay adapted from that described by Miller, In: *A Short Course in Bacterial Genetics* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY,) p. 72 (1992).

2. Results

Metabolite-dependent conformational changes in the 202-nucleotide leader sequence of the *btuB* mRNA. Figure 1A: Separation of spontaneous RNA-cleavage products of the *btuB* leader using denaturing 10% polyacrylamide gel electrophoresis (PAGE). 5'-32p-labeled mRNA leader molecules (arrow) were incubated for 41 hr at 25°C in 20 mM MgCl₂, 50 mM Tris-HCl (pH 8.3 at 25°C) in the presence (+) or absence (-) of 20 μ M of AdoCbl. Lanes containing RNAs that have undergone no reaction, partial digest with alkali, and partial digest with RNase T1 (G-specific cleavage) are identified by NR, OH, and T1, respectively. The location of product bands corresponding to cleavage after selected guanosine residues are identified by filled arrowheads. Light blue arrowheads labeled 1 through 8 identify eight of the nine

locations that exhibit effector-induced structure modulation, which experience an increase or decrease in the rate of spontaneous RNA cleavage. The image was generated using a phosphorimager (Molecular Dynamics), and cleavage yields were quantitated by using ImageQuant software. Figure 1B: Sequence and secondary-structure model for the 202-nucleotide leader sequence of *btuB* mRNA in the presence of AdoCbl. Putative base-paired elements are designated P1 through P9. Complementary nucleotides in the loops of P4 and P9 that have the potential to form a pseudoknot are juxtaposed. Nine specific sites of structure modulation are identified by light blue arrowheads. The asterisks demark the boundaries of the B₁₂ box (nucleotides 141-162). The coding region and the 38 nucleotides that reside immediately 5' of the start codon (nucleotides 241-243) were not included in the 202-nucleotide fragment. The 315-nucleotide fragment includes the 202-nucleotide fragment, the remaining 38 nucleotides of the leader sequence, and the first 75 nucleotides of the coding region.

The *btuB* mRNA leader forms a saturable binding site for AdoCbl. Figure 2A: The dependence of spontaneous cleavage of *btuB* mRNA leader on the concentration of AdoCbl effector as represented by site 1 (G23) and site 2 (U68). 5'-³²P-labeled mRNA leader molecules were incubated, separated, and analyzed as described in the legend to Figure 1A, and include identical control and marker lanes as indicated. Incubations contained concentrations of AdoCbl ranging from 10 nM to 100 μM (lanes 1 through 8) or did not include AdoCbl (-). Figure 2B: Composite plot of the fraction of RNA cleaved at six locations along the mRNA leader versus the logarithm of the concentration (c) of AdoCbl. Fraction cleaved values were normalized relative to the highest and lowest cleavage values measured for each location, including the values obtained upon incubation in the absence of AdoCbl. The inset defines the symbols used for each of six sites, while the remaining three sites were excluded from the analysis due to weak or obscured cleavage bands. Filled and open symbols represent increasing and decreasing cleavage yields, respectively, upon increasing the concentration of AdoCbl. The dashed line reflects a K_D of ~300 nM, as predicted by the concentration needed to generate half-maximal structural modulation. Data plotted were derived from a single PAGE analysis, of which two representative sections are depicted in Figure 2A.

The 202-nucleotide mRNA leader causes an unequal distribution of AdoCbl in an equilibrium dialysis apparatus. Figure 3(I): Equilibration of tritiated effector was conducted in the absence of RNA. Figure 3(II): (step 1) Equilibration was conducted as

in I, but with 200 pmoles of mRNA leader added to chamber b; (step 2) 5,000 pmoles of unlabeled AdoCbl was added to chamber b. Figure 3(III): Equilibrations were conducted as described in II, but wherein 5,000 pmoles of cyanocobalamin was added to chamber b. IV: (step 1) Equilibration was initiated as described in step 1 of II; (steps 2 and 3) the solution in chamber a was replaced with 25 μ L of fresh equilibration buffer; (step 4) 5,000 pmoles of unlabeled AdoCbl was added to chamber b. The cpm ratio is the ratio of counts detected in chamber b relative to that of a. The dashed line represents a cpm ratio of 1, which is expected if equal distribution of tritium is established.

Selective molecular recognition of effectors by the *btuB* mRNA leader. Figure 4A shows a chemical structure of AdoCbl (1) and various effector analogs (2 through 11). Figure 4B: Determination of analog binding by monitoring modulation of spontaneous cleavage of the 202-nucleotide *btuB* RNA leader. 5'-³²P-labeled mRNA leader molecules were incubated, separated, and analyzed as described in the legend to Figure 1A, and include identical control and marker lanes as indicated. The sections of three PAGE analyses encompassing site 2 (U68) are depicted. Below each image is plotted the amount of RNA cleaved (normalized with relation to the lowest and highest levels of cleavage at U68 in each gel) for each effector as indicated, or for no effector (-). The compound 11 (13-epi-AdoCbl) is an epimer of AdoCbl wherein the configuration at C13 is inverted, so that the ϵ propionamide side chain is above the plane of the corrin ring; see Brown et al., Conformational studies of 5'-deoxyadenosyl-13-epicobalamin, a coenzymatically active structural analog of coenzyme B₁₂. Polyhedron 17, 2213 (1998).

Mutations in the mRNA leader and their effects on AdoCbl binding and genetic control. Figure 5A: Sequence of the putative P5 element of the wild-type 202-nucleotide *btuB* leader exhibits AdoCbl-dependent modulation of structure as indicated by the observed increase in spontaneous RNA cleavage at position U68 (10% denaturing PAGE gel). Assays were conducted in the absence (-) or presence (+) of 5 μ M AdoCbl. The remaining lanes are as described in the legend to Figure 1A. The composite bar graph reflects the ability of the RNA to shift the equilibrium of AdoCbl in an equilibrium dialysis apparatus and the ability of a reporter gene (see Experimental Procedures) to be regulated by AdoCbl addition to a bacterial culture. (Left) Plotted is the cpm ratio derived by equilibrium dialysis, wherein chamber b contains the RNA. Details of the equilibrium dialysis experiments are described in the brief description of Figure 3. (Right) Plotted are the expression levels of β -galactosidase as determined from cells

grown in the absence (-) or presence (+) of 5 μ M AdoCbl. Boxed numbers on the left and right, respectively, reflect the approximate K_D and the fold repression of β -galactosidase activity in the presence of AdoCbl. N.D. designates not determined. Figures 5B-5F:

Sequences and performance characteristics of various mutant leader sequences as indicated. Constructs were created as described in the Experimental Procedures section.

i. Metabolite-induced structure modulation of a messenger RNA.

To assess whether the *btuB* leader sequence alone is sufficient for sensing and responding to a metabolite, a molecular probing strategy was employed that relies on the structure-dependent spontaneous cleavage of RNA (Soukup and Breaker, Relationship between internucleotide linkage geometry and the stability of P_i-NA. RNA 5, 1308-1325 (1999); Soukup et al., Generating new ligand-binding RNAs by affinity maturation and disintegration of allosteric ribozymes. RNA 7, 524-536 (2001)). The principal mechanism by which an RNA phosphodiester linkage is spontaneously cleaved involves an internal nucleophilic attack by the 2'-oxygen on the adjacent phosphorus center. Since the precise "in-line" positioning of the U-oxygen, phosphorus, and 5'-oxygen atoms of a given RNA linkage is essential for a productive nucleophilic attack to occur (Soukup and Breaker, Relationship between internucleotide linkage geometry and the stability of P_i-NA. RNA 5, 1308-1325 (1999); Soukup et al., Generating new ligand-binding RNAs by affinity maturation and disintegration of allosteric ribozymes. RNA 7, 524-536 (2001); Westheimer, Pseudo-rotation in the hydrolysis of phosphate esters. Acc. Chem. Res. 1, 70-78 (1968); Usher, On the mechanism of ribonuclease action. Proc. Natl. Acad. USA 62, 661-667 (1969); Usher and McHale, Hydrolytic stability of helical RNA: a selective advantage for the natural 3',5'-bond. Proc. Natl. Acad. USA 73, 1149-1153 (1976); Dock-Bregeon and Moras, Conformational changes and dynamics of tRNAs: evidence from hydrolysis patterns. Cold Spring Harbor Symp. Quant. Biol. 52, 113-121 (1987)), the rate at which spontaneous cleavage occurs at a given linkage is highly dependent upon the secondary and tertiary structure of the RNA. Specifically, RNA linkages that are formed by nucleotides involved in stable base-paired structures rarely undergo spontaneous cleavage because they rarely adopt an in-line conformation, while nucleotides located in relatively unstructured regions or in tertiary-structured regions experience far greater levels of spontaneous cleavage. Thus, probing of an RNA receptor in the absence and presence of its ligand can be used to provide evidence for RNA structural models and even to determine the dissociation constant for a given RNA-

ligand interaction (Soukup and Breaker, Relationship between internucleotide linkage geometry and the stability of P₁-NA. RNA 5, 1308-1325 (1999); Soukup et al., Generating new ligand-binding RNAs by affinity maturation and disintegration of allosteric ribozymes. RNA 7, 524-536 (2001)).

5 A preparation of RNAs that encompass nucleotides 1 through 202 of the 5'-untranslated region of the *btuB* mRNA (Nou and Kadner, Adenosylcobalamin inhibits ribosome binding to *btuB* RNA. Proc. Natl. Acad. Sci. USA 97, 7190-7195 (2000); Lundrigan et al., Transcribed sequences of the *Escherichia coli* *btuB* gene control its expression and regulation by vitamin B₁₂ Proc. Natl. Acad. Sci. USA 88, 1479-1483 (1991))
10 was subjected to in-line probing (Figure 1). In the absence of the putative AdoCbl effector, the RNA exhibits a distinct pattern of cleavage products that is indicative of a well ordered conformational state, which has a mixture of stable structural elements interspersed with regions that are mostly unstructured (Figure 1A). In the presence of AdoCbl, the pattern of cleavage changes at eight locations, while a ninth position of
15 structural modulation (Figure 1B) is observed when a longer portion of the mRNA is used. Specifically, metabolite-induced structural modulation at nucleotide 202 (Figure 1B, position 9) was observed by using in-line probing of a fragment that encompasses nucleotides 1 through 315 of the *btuB* mRNA (Nou and Kadner, Adenosylcobalamin inhibits ribosome binding to *btuB* RNA. Proc. Natl. Acad. Sci. USA 97, 7190-7195
20 (2000)). Positions 1, 3, 4, 8, and 9 undergo an effector-dependent dampening of spontaneous cleavage while the remaining sites experience the reverse effect. A similar pattern of metabolite-modulated RNA cleavage was observed with the analogous 206-nucleotide *btuB* leader RNA of *S. typhimurium* (Wei et al., Res. Microbiol. 143, 459 (1992)).

25 These effector-modulated sites are mapped on a secondary-structure model that was generated by using a combination of computational and RNA probing data. An RNA secondary-structure prediction algorithm (Zuker et al., Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide. In RNA Biochemistry and Biotechnology (eds. Barciszewski, J., and Clark, B. F. C.) pp. 11-43 (NATO ASI Series, Kluwer Academic Publishers) (1999)) supports a model wherein nine base-paired
30 elements are formed. The in-line probing data and preliminary mutational analyses are consistent with eight of these pairing interactions (P 1-P4 and P6-P9), while an alternative pairing interaction (P5) is supported (see below). The majority of these

putative base-paired elements appear to remain intact upon effector-induced modulation, with the notable exception of P9. The importance of this structural element in the modulation of ribosome binding and translation has been previously established by mutational analysis (Nou and Kadner, Adenosylcobalamin inhibits ribosome binding to *btuB* RNA. Proc. Natl. Acad. Sci. USA 97, 7190-7195 (2000)). Metabolite-dependent formation of the P9 stem-loop structure appears to be critical for the down-regulation of mRNA translation. Consistent with this hypothesis is the observed increase in structure formation in this location upon the addition of AdoCbl (Figure 1B, decreased cleavage at positions 8 and 9).

10 **ii. A saturable metabolite-binding site is formed by a messenger RNA.**

If the structural alteration of the mRNA leader is induced selectively by AdoCbl (as opposed to modulation by a non-specific effect) then the RNA should exhibit characteristics of a typical receptor-ligand interaction. Thus, a plot of the relative extents of structural modulation at each site is expected to yield an apparent dissociation constant (apparent K_D) for the effector, which reflects the concentration of effector needed to convert half of the RNAs into their altered structural state. Furthermore, if a single binding event brings about the global structural changes that are observed, then the individual K_r values calculated for each modulation site should converge on a single value, while these values are likely to vary if the structural modulation results from non-specific effects.

Indeed, the levels of spontaneous RNA cleavage were found to correlate with the concentrations of AdoCbl added to the in-line probing mixtures (Figure 2A). Examination of the dependency of the six most prominent sites of modulation on effector concentration reveals similar apparent K_D values of approximately 300 nM at 25°C (Figure 2B). This value is comparable to an apparent K_D value derived from a previous assay that examined the AdoCbl-dependent binding of ribosomes to the *btuB* mRNA (Nou and Kadner, Adenosylcobalamin inhibits ribosome binding to *btuB* RNA. Proc. Natl. Acad. Sci. USA 97, 7190-7195 (2000)). Moreover, the fact that structural modulation occurs over a broad range of concentrations of AdoCbl suggests that this RNA is not likely to make use of cooperative binding of multiple effectors, which would result in a more substantial response to small changes in effector concentration. Together, these observations indicate that the mRNA leader undergoes a substantial change in conformation and forms a high-affinity binding pocket for AdoCbl.

To provide further support for this conclusion, equilibrium dialysis was used to determine whether the RNA could selectively generate an unequal distribution of tritiated AdoCbl (3H-AdoCbl) when incubated in a two-chamber dialysis system. As expected, addition of 3H-AdoCbl to chamber a of an equilibrium dialysis assembly
5 results in near equal distribution of tritium (cpm ratio ~1) between chambers a and b upon incubation (Figure 3, experiment I). However, the addition of the 202-nucleotide mRNA leader to chamber b causes a shift in the equilibrium of 3H-AdoCbl (cpm ratio ~2) in favor of chamber b (Figure 3, experiments II and III). Importantly, the subsequent addition of an excess of unlabeled AdoCbl restores equal distribution of tritium between
10 the two chambers, while the addition of an excess of cyanocobalamin (vitamin B₁₂, an analog of AdoCbl) does not restore the ratio of tritium to unity. Excess unlabeled AdoCbl is expected to restore equal distribution by serving to occupy the vast majority of the binding sites formed by the *btuB* RNA. In contrast, cyanocobalamin is known to be incapable of serving as a regulatory effector for *btuB* expression in *E. coli* (Nou and
15 Kadner, Adenosylcobalamin inhibits ribosome binding to *btuB* RNA. Proc. Natl. Acad. Sci. USA 97, 7190-7195 (2000); Lundrigan and Kadner, Altered cobalamin metabolism in *Escherichia coli* *btuR* mutants affects *btuB* gene regulation. J. Bacteriol. 171, 154-161 (1989)), and thus should be ignored as an effector by the RNA. These findings are consistent with the conclusion that the RNA directly binds AdoCbl and indicate that the
20 RNA forms a selective binding pocket that excludes certain analog compounds.

Assuming that a 1:1 complex is formed between effector and RNA, it was expected that equilibrium dialysis would produce a cpm ratio of far greater than 2 under the assay conditions (2-fold excess RNA over 3H-AdoCbl and concentrations of RNA and effector in excess of the apparent K_D). Since there should be an excess of binding
25 sites, the majority of the tritium should be shifted to chamber b upon equilibration. However, the data suggest that ~70% of the tritium in the sample used is not in the form of 3H-AdoCbl. For example, successive replacement of the buffer in chamber a (which removes unshifted tritium from the equilibrium dialysis system) results in increasing values for the cpm ratio (Figure 3; experiment IV). In addition, the tritium that remains
30 in chamber a upon equilibration with RNA in chamber b cannot be induced to yield an unequal distribution of tritium by *btuB* RNA in a subsequent equilibrium dialysis experiment (data not shown). The source of this unbound tritium is most likely from light-mediated degradation of AdoCbl, which is highly unstable under ambient light

conditions. Mass spectrum analysis of ^3H -AdoCbl reveals that the sample is almost entirely intact in the absence of light exposure, but yields ~70% degradation upon exposure to light for a time of about 20 sec) that is typically experienced by a sample when establishing an equilibrium dialysis experiment.

5 **iii. The *btuB* mRNA leader selectively binds AdoCbl.**

To provide selectivity for the genetic response, the *btuB* mRNA leader must form a precise binding pocket for AdoCbl in order to preclude the genetic switch from being triggered by other metabolites. To explore the molecular recognition capabilities of this RNA, the binding affinity of AdoCbl relative to 10 analogs was indirectly determined (Figure 4A). This was achieved by determining the extent of spontaneous cleavage at site 2 (nucleotide U68) upon incubation in the presence of AdoCbl or of various analogs (Figure 4B). It was found that the RNA fails to undergo structural modulation when cobalamin compounds lack the 5'-deoxy-5'-adenosyl moiety. The importance of individual functional groups on this moiety is revealed by the function of other analogs. In summary, modifications at the N1, N3, and N6 positions of the adenine ring cause significant disruption of binding, while the 2"-hydroxyl group of the adjoining ribose moiety is not an important molecular recognition element. Interestingly, a change in the stereochemistry at position 13 of the corrin ring (compound 11) renders the molecule inactive as a regulatory effector in this *in vitro* assay and also inside cells. These findings indicate that the *btuB* mRNA leader forms a binding pocket for AdoCbl and that the RNA makes numerous contacts with the effector to ensure high molecular specificity.

15 **iv. Disruption of metabolite-RNA binding has consequences for genetic control.**

The presence of AdoCbl causes reductions in ribosome binding and translation efficiency of the *btuB* mRNA (Nou and Kadner, Adenosylcobalamin inhibits ribosome binding to *btuB* RNA. Proc. Natl. Acad. Sci. USA 97, 7190-7195 (2000)). The results indicate that this genetic control process is mediated by the selective binding of AdoCbl to the *btuB* mRNA. The effector-binding function of mutant RNA leaders *in vitro* was compared with their ability to support effector-induced genetic control inside cells. As expected, the wild-type mRNA leader exhibits effector-induced structure modulation, induces an unequal distribution of ^3H -AdoCbl in an equilibrium dialysis system, and permits down regulation of a reporter gene in *E. coli* cells treated with AdoCbl and harboring the appropriate reporter construct (summarized in Figure 5A). However, the

introduction of a single mutation (A150T) in the evolutionarily conserved "B₁₂ box" (Nou and Kadner, Adenosylcobalamin inhibits ribosome binding to *btuB* RNA. Proc. Natl. Acad. Sci. USA 97, 7190-7195 (2000)) completely eliminates the *in vitro* effector-binding and *in vivo* gene-control functions of this construct, termed "m1" (Figure 5B), which is consistent with the necessity of effector binding for genetic control.

Mutations that disrupt (U73G, G74U) and subsequently restore (U73G, G74U, C114A, A115C) the predicted P5 stem element were examined. The disrupted stem in construct m2 causes a reduction of AdoCbl binding affinity *in vitro* and a corresponding reduction of genetic control in cell assays (Figure 5C), while restoration of the P5 stem element (construct m3) results in near wild-type functions for binding and genetic control (Figure 5D). This indicates that the P5 stem is an important structural element for function of the RNA. Interestingly, potentially disruptive (m4) and restorative (m5) mutations in a possible pseudoknot structure between the P4 and P9 loops (Figure 1B) both result in a reduction in binding affinity ($K_D \sim 5 \mu\text{M}$). If a pseudoknot is being formed, this structure might require a specific sequence for proper function. Although these RNAs maintain diminished but detectable levels of effector binding, neither exhibits genetic control upon the addition of AdoCbl to bacterial cultures harboring the corresponding reporter constructs. The loss in binding affinity likely is sufficient to place these mutant RNAs out of the physiological range for effector concentration, as the cells still retain their natural *btuB* gene whose regulatory system continues to control the import of AdoCbl. The findings support the hypothesis that mRNAs have the structural and functional sophistication needed to perform precision genetic control in the absence of protein regulatory elements.

v. Analysis

Genetic control by mRNAs that directly sense the concentrations of metabolites is a newly established paradigm for monitoring the status of cellular metabolism. Although sensing of aminoacyl tRNAs in prokaryotes also appears to be achieved by direct binding of tRNAs to the 5'-untranslated region of their corresponding aminoacyl tRNA synthetases (Henkin, tRNA-directed transcription antitermination. Mol. Microbiol. 3, 381-387 (1994)), binding appears to be mediated by Watson/Crick base pairing. In the case of *btuB* the mRNA directly binds the Ado-Cbl effector and becomes resistant to translation initiation, presumably by preventing ribosome binding (Nou and Kadner, Adenosylcobalamin inhibits ribosome binding to *btuB* RNA. Proc. Natl. Acad. Sci. USA

97, 7190-7195 (2000)). If no protein receptors are required for molecular recognition or for modulating gene expression, then this simple "riboswitch" mechanism is most economical in its architecture. Given the organizational simplicity of the *btuB* genetic control components compared to analogous systems that involve proteins, it is likely that mRNAs could be more easily engineered to respond directly to natural and non-biological regulatory effectors.

It is possible that variations of this mechanism involving direct contacts between metabolite and mRNA are far more widespread in genetic circuitry. For example, the *S. typhimurium cob* operon, which encodes proteins in the biosynthetic pathway for the AdoCbl coenzyme, carries B₁₂ box and other regulatory structures in its leader domain (Ravnum and Andersson, An adenosyl-cobalamin (coenzyme-B₁₂)-repressed translational enhancer in the *cob* mRNA of *Salmonella typhimurium*. Mol. Microbiol. 39, 1585-1594 (2001)). It has been noted (White III, Coenzymes as fossils of an earlier metabolic state. J. Mol. Evol. 7, 101-104 (1976)) that these two coenzymes and FMN, which is another potential riboswitch effector (Gelfand et al., A conserved RNA structure element involved in the regulation of bacterial riboflavin synthesis genes. Trends Genetics 15, 439-442 (1999)), possibly are molecular fossils of an ancient metabolic state that was run entirely by RNA. If true, then mechanisms involving metabolite sensing by mRNA might be one of the oldest forms of genetic control in existence.

B. Example 2: Thiamine Pyrophosphate (TPP) Riboswitches

The example described testing and analysis of a riboswitch that controls gene expression by binding thiamine pyrophosphate.

1. Chemicals and oligonucleotides

TPP, thiamine monophosphate (TP), thiamine, oxythiamine, amprolium, and benfotiamine were purchased from Sigma. Thiamine disulfide and 4-methyl-5-β-hydroxyethylthiazole (THZ) were purchased from TCI America. ³H-labeled thiamine was purchased from American Radiolabeled Chemicals, Inc. (10 Ci mmol⁻¹). Synthetic DNAs were synthesized by the Keck Foundation Biotechnology Resource Center at Yale University. DNAs were purified by denaturing (8 M urea) polyacrylamide gel electrophoresis (PAGE) and isolated from the gel by crush-soaking in 10 mM Tris-HCl

(pH 7.5 at 23°C), 200 mM NaCl and 1 mM EDTA. The DNA was recovered by precipitation with ethanol.

2. Construction of *E. coli* *thiM*- and *E. coli* *thiC*-*lacZ* fusions

Nucleotides -83 to 238 of the *E. coli thiCEFGH* operon (Vander Horn et al.,
5 Structural genes for thiamine biosynthetic enzymes (*thiCEFGH*) in *E. coli* K-12.
J. Bacteriology **175**, 982-992 (1993)), was amplified by PCR from *E. coli* strain MC4100
(obtained from S. Gottesman, NIH) as a *EcoRI*-*Bgl* II fragment. The DNA was ligated
into *EcoRI*- and *BamHI*-digested pRS414 plasmid DNA, which contains a promoterless
copy of *lacZ* (obtained from R. Simons, UCLA; Simons et al., Improved single and
10 multicopy lac-based cloning vectors for protein and operon fusions *Gene* **53**, 85-96
(1987)), resulting in the in-frame fusion of the 9th codon of *lacZ* to the 9th codon of *thiC*.
Similarly, the regulatory region of *thiM* (nucleotides -67 to 163) was amplified by PCR
as a *EcoRI*-*BamHI* fragment and inserted into plasmid pRS414, wherein the 6th codon of
thiM resides in-frame with the 9th codon of *lacZ*. The plasmids were transformed into
15 Top10 cells (Invitrogen) for all subsequent manipulations. All site-directed mutations
were introduced into the *thiC* and *thiM* regulatory regions using the QuikChange site-
directed mutagenesis kit (Stratagene) and the appropriate mutagenic DNA primers. All
mutations were confirmed by DNA sequencing (USB Thermosequenase).

3. Thiamine-repression β -galactosidase assays

20 *E. coli* cells (Top10; Invitrogen) that contained in-frame *lacZ* fusions to *thiC* or
thiM mRNA leader sequences, were grown in M9 glucose minimal media (plus 50 μ g/ml
Vitamin assay Casamino acids; Difco) to mid-exponential phase. The cultures were
either grown with or without added thiamine (100 μ M). Aliquots (1 mL) were removed
for β -galactosidase enzyme assays, which were conducted in a manner similar to that
25 described by Miller (Miller, In: *A Short Course in Bacterial Genetics* Cold Spring
Harbor Laboratory Press, Cold Spring Harbor, NY, p. 72. (1992)). All assays were
repeated twice and in duplicate, with Miller unit values reflecting the average of these
analyses.

4. *In vitro* transcription

30 Templates for *in vitro* transcription of the fragments of *thiC* and *thiM* mRNA
leaders were generated by PCR using the appropriate DNA primers and plasmids
pRS414*thiC* or pRS414*thiM*, respectively. The dinucleotide sequence GG was

introduced into the DNA constructs (corresponding to the 5' terminus of each RNA construct) at this step to facilitate transcription by T7 RNA polymerase. RNAs were prepared by *in vitro* transcription and were 5' ³²P-labeled as described previously (Seetharaman et al., Immobilized riboswitches for the analysis of complex chemical and biological mixtures. *Nature Biotechnol.* 19, 336-341 (2001)).

5. In-line probing of RNA

Determination of apparent K_D values for each construct was achieved by conducting in-line probing of RNA constructs wherein the concentration of the ligand was varied between 10 nM and 100 μ M, or up to 10 mM for weakly binding ligands. Specifically, TPP-dependent modulation of the spontaneous cleavage of RNA constructs was visualized by polyacrylamide gel electrophoresis (PAGE). 5' ³²P-labeled RNAs (20 nM) were incubated for approximately 40 hr at 25°C in 20 mM MgCl₂, 50 mM Tris-HCl (pH 8.3 at 25°C) in the presence (+) or absence (-) of 100 μ M TPP. Some RNAs were subjected to no reaction, partial digestion with alkali, or partial digestion with RNase T1 (G-specific cleavage) (see Figure 6a). Composite plots of the fraction of RNA cleaved at specific sites versus the logarithm of the concentration of ligand (e.g. Figure 7a) were generated to provide an estimate of the apparent K_D . Fraction cleaved values were normalized relative to the highest and lowest cleavage values measured for each site.

6. Equilibrium dialysis

Equilibrium dialysis assays were conducted using a DispoEquilibrium Dialyzer (ED-1, Harvard Bioscience), wherein chambers *a* and *b* were separated by a 5,000 Dalton molecular weight cut-off membrane. Equilibration was initiated by the addition of 25 μ L of equilibration buffer [50 mM Tris-HCl (pH 8.3 at 25°C), 20 mM MgCl₂, 100 mM KCl], containing 100 nM ³H-thiamine and by the addition of an equal volume of equilibration buffer either without or with 20 μ M RNA as indicated to chamber *b*. Equilibrations were allowed to proceed for 10 hr at 23°C, and aliquots were removed from each chamber and quantitated by using a liquid scintillation counter.

7. Results

i. Metabolite binding by mRNAs.

Figure 6A shows TPP-dependent modulation of the spontaneous cleavage of 16S *thiM* RNA was visualized by polyacrylamide gel electrophoresis (PAGE). 5' ³²P-labeled RNAs (arrow, 20 nM) were incubated for approximately 40 hr at 25°C in 20 mM MgCl₂,

50 mM Tris-HCl (pH 8.3 at 25°C) in the presence (+) or absence (–) of 100 μ M TPP. NR, [–]OH and T1 represent RNAs subjected to no reaction, partial digestion with alkali, or partial digestion with RNase T1 (G-specific cleavage), respectively. Product bands representing cleavage after selected G residues are numbered and identified by filled
 5 arrowheads. The asterisk identifies modulation of RNA structure involving the Shine-Dalgarno (SD) sequence. Gel separations were analyzed using a phosphorimager (Molecular Dynamics) and quantitated using ImageQuant software.

Figure 6B shows a secondary-structure model of 165 *thiM* as predicted by computer modeling (Zuker et al., Algorithms and thermodynamics for RNA secondary
 10 structure prediction: a practical guide. In *RNA Biochemistry and Biotechnology* (eds. Barciszewski J. & Clark, B.F.C.) 11-43 (NATO ASI Series, Kluwer Academic Publishers, 1999); Mathews et al., Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* **288**, 911-940 (1999)) and by the structure probing data depicted in Figure 6A. Spontaneous cleavage
 15 characteristics are as noted in the inset. Unmarked nucleotides exhibit a constant but low level of degradation. The truncated 91 *thiM* RNA is boxed and the *thi* box element (Miranda-Rios et al., A conserved RNA structure (*thi* box) is involved in regulation of thiamin biosynthetic gene expression in bacteria. *Proc. Natl. Acad. Sci. USA* **98**, 9736-9741 (2001)) is shaded light blue. Nucleotides highlighted in orange identify an
 20 alternative pairing, designated P8*. The RNA carries two mutations (G156A and U157C) relative to wild type that were introduced in a non-essential portion of the construct to form a restriction site for cloning, while all RNAs carry two 5'-terminal G residues to facilitate *in vitro* transcription.

Figure 6C shows TPP-dependent modulation of the spontaneous cleavage of 240
 25 *thiC* RNA. Reactions were conducted and analyzed as described in above for Figure 6A. Figure 6D shows a secondary-structure model of 240 *thiC*. Base-paired elements that are similar to those of *thiM* are labeled P1 through P5. The truncated RNA 111 *thiC* is boxed. Nucleotides highlighted in orange identify an alternative pairing.

**ii. The *thiM* and *thiC* mRNA leaders serve as high-affinity metabolite
 30 receptors.**

Figure 7A shows the extent of spontaneous modulation of RNA cleavage at several sites within 165 *thiM* (left) and 240 *thiC* (right) plotted for different concentrations (*c*) of TPP. Red arrows reflect the estimated concentration of TPP needed

to attain half maximal modulation of RNA (apparent K_D). Figure 7B shows the logarithm of the apparent K_D values plotted for both RNAs with TPP, TP and thiamine as indicated. The boxed data was generated using TPP with the truncated RNAs 91 *thiM* and 111 *thiC*. Figure 7C shows that patterns of spontaneous cleavage of 165 *thiM* differ between thiamine and TPP ligands as depicted by PAGE analysis (left) and as reflected by graphs (right) representing the relative phosphorimager counts for the three lanes as indicated. Details for the RNA probing analysis are similar to those described above in connection with Figure 6A. The graphs were generated by ImageQuant software.

iii. High sensitivity and selectivity of mRNA leaders for metabolite binding.

Figure 8A shows chemical structures of several analogues of thiamine. TD is thiamine disulfide and THZ is 4-methyl-5- β -hydroxyethylthiazole. Figure 8B shows PAGE analysis of 165 *thiM* RNA structure probing using TPP and various chemical analogues (40 μ M each) as indicated. Locations of significant structural modulation within the RNA spanning nucleotides ~113 to ~150 are indicated by open arrowheads. The asterisk identifies the site (C144) used to compare the normalized fraction of RNA that is cleaved (bottom) in the presence of specific compounds. Details for the RNA probing analysis are similar to those described above in connection with Figure 6A. Figure 8C shows a summary of the features of TPP that are critical for molecular recognition. Figure 8D shows equilibrium dialysis using 3 H-thiamine as a tracer. Plotted are the ratios for tritium distribution in a two-chamber system (*a* and *b*) that were established upon equilibration in the presence of the RNA constructs in chamber *b* as indicated (see below for a description of the non-TPP-binding mutant M3). 100 μ M TPP or oxythiamine were added to chamber *a*, as denoted, upon the start of equilibration.

iv. Mutational analysis of the structure and function of the *thiM* riboswitch.

Figure 9A shows mutations present in constructs M1 through M8 relative to the 165 *thiM* RNA. P8* is a putative base-paired element between portions (orange) of the P1 and P8 stems. Figure 9B (top) shows *in vitro* ligand-binding and genetic control functions of the wild-type (WT), M1 and M2 RNAs as reflected by PAGE analysis of in-line probing experiments (10 μ M TPP) and by β -galactosidase expression assays. Labels on PAGE gels are as described above in connection with Figure 6A. Bars represent the levels of gene expression in the presence (+) and the absence (-) of TPP in the culture medium. Figure 9C is a summary of similar analyses of WT through M9 is presented in

table form. The SD status "n.d." (not determined) indicates either that the level of spontaneous cleavage detected in the absence and presence of TPP is near the limit of detection (M6, M7 and M8) or that the region adopts an atypical structure (M9) compared to WT.

5 8. Discussion

β -galactosidase fusion constructs were prepared that encompass the 5'-untranslated region of *thiM* and *thiC* mRNAs of *E. coli*, which includes a previously identified "*thi* box" domain whose sequence and potential secondary structure are conserved in several species of bacteria and archaea (Miranda-Rios et al., A conserved RNA structure (*thi* box) is involved in regulation of thiamin biosynthetic gene expression in bacteria. *Proc. Natl. Acad. Sci. USA* 98, 9736-9741 (2001)). The *thiM* and *thiC* translational fusion constructs exhibit thiamine-dependent suppression of β -galactosidase activity of 18- and 110-fold, respectively, when host cells are grown in a minimal medium that otherwise lacks a source of thiamine. A transcriptional fusion containing the *thiM* leader is not subject to suppression by thiamine, but a similar fusion with *thiC* leader yields a 16-fold modulation with thiamine, suggesting that a significant portion of genetic control observed with *thiC* occurs at the level of transcription.

 These constructs were used to prepare DNA templates by PCR for *in vitro* transcription of RNA fragments. The resulting RNAs were subjected to a structure-probing process (see Example 1) to reveal whether the RNAs undergo structure modulation upon binding of ligands. Internucleotide linkages in unstructured regions are more likely to undergo spontaneous cleavage compared to linkages that reside in highly structured regions of an RNA (Soukup & Breaker, Relationship between internucleotide linkage geometry and the stability of RNA. *RNA* 5, 1308-1325 (1999)). The 165-nucleotide *thiM* RNA fragment (165 *thiM*) has a distinct pattern of cleavage products that is generated when the RNA is incubated for an extended period in the absence of TPP (Figure 6A). Upon addition of 100 μ M TPP, 165 *thiM* undergoes substantial structural alteration as many internucleotide linkages in the region spanning positions 39 through 80 exhibit a reduction in spontaneous cleavage. This indicates that TPP binds to the RNA and stabilizes a defined structure within this region, resulting in a lower rate of fragmentation.

The fragmentation patterns are largely congruent with potential base-paired and bulge structures that are identified by a secondary-structure prediction algorithm (Zuker et al., Algorithms and thermodynamics for RNA secondary structure prediction: a practical guide. In *RNA Biochemistry and Biotechnology* (eds. Barciszewski J. & Clark, B.F.C.) 11-43 (NATO ASI Series, Kluwer Academic Publishers, 1999); Mathews et al., Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* **288**, 911-940 (1999)). Most linkages that experience a ligand-induced reduction of cleavage are encompassed by the *thi* box and nucleotides that reside immediately 5' relative to this domain (Figure 6B). Other linkages that undergo cleavage, but that are not modulated by TPP, are predicted to reside in bulges or in the loops of hairpins. Predicted base-paired structures labeled P2 through P7 encompass linkages that exhibit the lowest levels of spontaneous cleavage, implying that they remain structured in both the presence and absence of TPP. Interestingly, nucleotides 126 through 130 encompass the only region apart from those described above that become more structured upon TPP addition. These nucleotides correspond to the Shine-Dalgarno (SD) sequence, which is required for efficient translation of mRNAs in prokaryotes. These findings are consistent with a genetic control mechanism wherein the *thiM* RNA binds to TPP and forms a complex wherein the ribosome cannot gain access to the SD sequence.

Similarly, structure probing was used to examine the mRNA leader for *thiC*. The 240 *thiC* RNA also exhibits extensive modulation of its pattern of spontaneous cleavage, and again the majority of the changing pattern is located in the *thi* box and in the region located immediately upstream of this domain (Figure 6C). These regions of highest structure modulation in *thiM* and *thiC* can be folded into similar secondary structures (Figure 6D), and carry several common sequence elements within and adjacent to the *thi* box domain. Thus, the structures of *thiM* and *thiC* spanning stems P1 through P5 comprise TPP-binding motifs that are analogous to aptamers, which are engineered ligand-binding RNAs (Osborne & Ellington, Nucleic acid selection and the challenge of combinatorial chemistry. *Chem. Rev.* **97**, 349-370 (1997); Hermann & Patel, Adaptive recognition by nucleic acid aptamers. *Science* **287**, 820-825 (2000); Gold et al., Diversity of oligonucleotide functions. *Annu. Rev. Biochem.* **64**, 763-797 (1995)). Nucleotides residing 3' relative to this natural TPP aptamer are involved in converting the metabolite binding event into a genetic response.

The sensitivity of metabolite detection by these mRNAs was assessed by establishing apparent dissociation constant (apparent K_D) values for TPP, thiamine, and thiamine monophosphate (TP). Values were generated by monitoring the extent of spontaneous cleavage at several ligand-sensitive sites within the RNA under a range of ligand concentrations. For example, probing of a trace amount of 165 *thiM* RNA under TPP concentrations ranging from zero to 100 μ M (or up to 10 mM with certain analogues) reveals that half-maximal modulation of RNA structure occurs when approximately 600 nM TPP is present (Figure 7A), which reflects an apparent K_D of 600 nM. Likewise, probing of 240 *thiC* reveals an apparent K_D of 100 nM. Both 165 *thiM* and 240 *thiC* RNAs appear to bind TPP more avidly than TP or thiamine, with *thiC* exhibiting more than 1,000-fold discrimination against TP and thiamine (Figure 7B). The fact that TPP is the strongest modulator of RNA structure is consistent with genetic observations in *Salmonella typhimurium* that TPP synthesis is required for regulation of expression of thiamine biosynthesis genes (Webb et al., Thiamine pyrophosphate (TPP) negatively regulates transcription of some *thi* genes of *Salmonella typhimurium*. *J. Bacteriol.* 178, 2533-2538 (1996)). The differential specificity achieved by the RNAs, which is a phenomenon that is commonly observed for receptor-ligand systems made of protein, indicates that these ligand-binding RNAs would be receptive to specificity changes (through, for example, natural or artificial evolutionary forces).

The actual K_D values for RNA-ligand interactions might be different inside cells where physiological conditions of Mg^{2+} and other agents that can influence RNA structure will not match those of the *in vitro* assays. Also, the nature of the RNA construct can be a source of an altered K_D . For example, the minimized 91 *thiM* construct (Figure 6A), which largely encompasses only the putative natural aptamer, retains the ability to bind TPP and exhibits an apparent K_D that is improved by approximately 20 fold compared to the longer construct (Figure 7B). Thus, the affinity for TPP might vary as the nascent RNA transcript emerges from the active site of RNA polymerase or the ribosome. Furthermore, this result demonstrates that the 91 *thiM* aptamer domain can be separated from RNA components (collectively termed the “expression platform”) that are directly controlling gene expression. This modular construction, involving the physical and functional separation of aptamer and expression platform domains allows the generation of TPP-controlled RNAs by rational RNA engineering strategies (or through evolutionary processes).

Spontaneous cleavage at several linkages within the *thi* box domain of 165 *thiM* specifically correlate with the type of ligand used. Although TPP reduces spontaneous cleavage of 165 *thiM* at nucleotides A61, U62 and to a smaller extent at U79, these same sites retain an elevated level of cleavage when thiamine is present near its saturating concentration (Figure 7C). These nucleotides cluster at an internal bulge within the *thi* box domain, and appear to contribute to the binding site for the phosphate groups of TPP.

The structural modulation of 165 *thiM* was further examined in the presence of several analogues that carry certain structural features of thiamine (Figure 8A). Thiamine and its phosphorylated derivatives TP and TPP induce modulation as expected (Figure 8B). However, oxythiamine and other thiamine analogues with less similarity to TPP fail to induce structure modulation. The performance of this sampling of analogues indicates that the RNA makes specific contacts to distal parts of its ligand and that both the purine and phosphate groups carry important elements for molecular recognition (Figure 8C). Similar results are obtained by using equilibrium dialysis assays (Figure 8D). For example, the addition of 91 *thiM* RNA to chamber *b* of an equilibrium dialysis assembly causes a shift in the distribution of ³H-thiamine in favor of chamber *b*, unless an excess of unlabeled TPP is also included. However, the presence of oxythiamine does not significantly restore the tritium distribution to unity, which is expected because probing data indicate that it is not able to bind the RNA. These findings indicate that the aptamer domain of the TPP riboswitch is highly selective for its target ligand.

The secondary structure model for 165 *thiM* RNA was examined in greater detail by generating and testing a series of variant constructs (Figure 9A). For example, variant M1 carries a mutation that disrupts the predicted P3 pairing element. This mutation causes a loss of TPP binding (Figure 9B, e.g. see position C77) and a loss of genetic control of the corresponding β -galactosidase fusion construct (Figure 9B, graph). Re-establishment of base pairing in the double-mutant construct M2 restores both TPP binding and genetic control. Similarly, disruptive and restorative mutations encompassed by constructs M3 through M6 are consistent with the formation of stems P5 and P8. Upon the addition of TPP, the SD element of both the WT and M2 constructs becomes sequestered in a structure that precludes a high level of spontaneous cleavage. In contrast, the M1 construct does not exhibit SD modulation (Figure 9B, nucleotides

126-130). These results are consistent with the genetic switch being turned off by a mechanism whereby TPP binding ultimately promotes the stable formation of P8, which reduces access to the SD by the ribosome.

5 The partner of the SD sequence in P8 (nucleotides 108 to 111) remains resistant to spontaneous cleavage both in the presence and absence of TPP (Figure 6A). This is consistent with the formation of P8, upon addition of TPP, due to the displacement of an alternative structure that otherwise prevents this anti-SD element from forming P8. Furthermore, nucleotides 83 through 86 are complementary to the anti-SD element and this region also resists spontaneous cleavage in the presence and absence of TPP. A
10 mechanism by which genetic control could result, which is tested as described below, is via the mutually exclusive formation of P8* in the 'On' state versus the simultaneous formation of P1 and P8 in the metabolite-bound 'Off' state (Figure 9C).

Constructs M7 through M9 were tested in an assessment of this mechanism. Construct M7 carries a U109C mutation in the anti-SD sequence that is designed to
15 destabilize the P8 interaction while simultaneously destabilizing the P8* interaction. M7 retains TPP binding function and exhibits a significant level of genetic modulation (Figure 9B, box), which is expected if the mutation does not disrupt the relative distribution of mRNAs between the 'On' and 'Off' states. In comparison, M8 (U110C) retains TPP binding, exhibits a dramatic reduction in the level of reporter expression, and
20 loses nearly all genetic modulation. In addition, M8 no longer exhibits detectable spontaneous cleavage in the SD sequence, which is consistent with the thermodynamic balance between P8 and P8* formation being shifted decidedly in favor of P8 in this RNA variant. Construct M9, which carries four mutations in the anti-SD element, has a significantly different pattern of spontaneous cleavage in the SD region. M9 fails to
25 reduce gene expression upon thiamine addition to cells, despite the fact that the construct retains TPP binding *in vitro*. It is evident from these data that TPP binding restricts the structural freedom of the SD element in the appropriate RNA variants, and that this correlates with genetic control.

C. Example 3: Metabolite-binding Riboswitches

30 1. Introduction

Modern organisms must coordinate the expression of many hundreds of genes in response to metabolic demands and environmental changes. Each gene product must be regulated temporally, quantitatively, and oftentimes spatially. Additionally, genetic

control processes must be dynamic, rapid, and selectively responsive to the specific conditions undergoing change. Therefore, organisms require sentries of genetic regulatory factors that continuously quantify a multitude of environmental signals. Upon measurement of a particular signal, which may be one of many possible biochemical or physical cues, these regulatory factors must modulate expression of a specific subset of the organism's genes.

It has generally been assumed that proteins are the obligate sensors of these signals because proteins are a proven medium for forming highly responsive sensors. However, it was discovered that mRNAs also are capable of acting as direct sensors of chemical and physical conditions for the purpose of genetic control. Classes of mRNA domains, collectively referred to as 'riboswitches', serve as RNA genetic control elements that sense the concentrations of specific metabolites by directly binding the target compound. Riboswitches that have been discovered are responsible for sensing metabolites that are critical for fundamental biochemical processes including adenosylcobalamin (AdoCbl) (see Example 1), thiamine pyrophosphate (TPP) (see Example 2), flavin mononucleotide (FMN), *S*-adenosylmethionine (SAM) (see Example 7), lysine (see Example 5), guanine (see Example 6), and adenine (see Example 8). Upon interaction with the appropriate small molecule ligand, riboswitch mRNAs undergo a structural reorganization that results in the modulation of genes that they encode. To date, all riboswitches that have been examined in detail cause genetic repression upon binding their target ligand, although riboswitches that activate gene expression upon ligand binding can be produced (and will likely be found in nature).

In each instance, riboswitch domains have been subjected to a battery of biochemical and genetic analyses in order to convincingly demonstrate that direct interaction of small organic metabolites with mRNA receptors leads to a corresponding alteration in genetic expression. This example provides a brief summary of these efforts and of some of the general characteristics that are exhibited by riboswitches. Using these discoveries and the principles of riboswitch operation described in this example and elsewhere herein, those of skill in the art can use and adapt riboswitches for many purposes including use as genetic tools and as targets for development of antimicrobials.

2. General Organization of Riboswitch RNAs

Bacterial riboswitch RNAs are genetic control elements that are located primarily within the 5'-untranslated region (5'-UTR) of the main coding region of a particular

mRNA. Structural probing studies (discussed further below) revealed that riboswitch elements are generally composed of two domains: a natural aptamer (T. Hermann, D. J. Patel, *Science* 2000, 287, 820; L. Gold, et al., *Annual Review of Biochemistry* 1995, 64, 763) that serves as the ligand-binding domain (referred to herein as the aptamer domain),
5 and an 'expression platform' that interfaces with RNA elements that are involved in gene expression (e.g. Shine-Dalgarno (SD) elements; transcription terminator stems). These conclusions are drawn from the observation that aptamer domains synthesized *in vitro* bind the appropriate ligand in the absence of the expression platform (see Examples 2 and 6). Moreover, structural probing investigations suggest that the aptamer domain of
10 most riboswitches adopts a particular secondary- and tertiary-structure fold when examined independently, that is essentially identical to the aptamer structure when examined in the context of the entire 5' leader RNA. This implies that, in many cases, the aptamer domain is a modular unit that folds independently of the expression platform (see Examples 2 and 6).

15 Ultimately, the ligand-bound or unbound status of the aptamer domain is interpreted through the expression platform, which is responsible for exerting an influence upon gene expression. The view of a riboswitch as a modular element is further supported by the fact that aptamer domains are highly conserved amongst various organisms (and even between kingdoms as is observed for the TPP riboswitch, whereas
20 the expression platform varies in sequence, structure, and in the mechanism by which expression of the appended open reading frame is controlled. For example, ligand binding to the TPP riboswitch of the *tenA* mRNA of *B. subtilis* causes transcription termination. This expression platform is distinct in sequence and structure compared to the expression platform of the TPP riboswitch in the *thiM* mRNA from *E. coli*, wherein
25 TPP binding causes inhibition of translation by a SD blocking mechanism (see Example 2). The TPP aptamer domain is easily recognizable and of near identical functional character between these two transcriptional units, but the genetic control mechanisms and the expression platforms that carry them out are very different.

Aptamer domains for riboswitch RNAs typically range from ~70 to 170 nt in
30 length (Figure 11). This observation was somewhat unexpected given that *in vitro* evolution experiments identified a wide variety of small molecule-binding aptamers, which are considerably shorter in length and structural intricacy (T. Hermann, D. J. Patel, *Science* 2000, 287, 820; L. Gold, et al., *Annual Review of Biochemistry* 1995, 64,
105

763; M. Famulok, *Current Opinion in Structural Biology* 1999, 9, 324). The substantial increase in complexity and information content of the natural aptamer sequences relative to artificial aptamers is most likely required to form RNA receptors that function with high affinity and selectivity. Apparent K_D values for the ligand-riboswitch complexes range from low nanomolar to low micromolar. It is also worth noting that some aptamer domains, when isolated from the appended expression platform, exhibit improved affinity for the target ligand over that of the intact riboswitch (~10 to 100-fold) (see Example 2). This likely represents an energetic cost in sampling the multiple distinct RNA conformations required by a fully intact riboswitch RNA, which is reflected by a loss in ligand affinity. Since the aptamer domain must serve as a molecular switch, this might also add to the functional demands on natural aptamers that might help rationalize their more sophisticated structures.

3. Riboswitch Regulation of Transcription Termination in Bacteria

Bacteria primarily make use of two methods for termination of transcription. Certain genes incorporate a termination signal that is dependent upon the Rho protein (J. P. Richardson, *Biochimica et Biophysica Acta* 2002, 1577, 251), while others make use of Rho-independent terminators (intrinsic terminators) to destabilize the transcription elongation complex (I. Gusarov, E. Nudler, *Molecular Cell* 1999, 3, 495; E. Nudler, M. E. Gottesman, *Genes to Cells* 2002, 7, 755). The latter RNA elements are composed of a GC-rich stem-loop followed by a stretch of 6-9 uridyl residues. Intrinsic terminators are widespread throughout bacterial genomes (F. Lillo, et al., *Bioinformatics* 2002, 18, 971), and are typically located at the 3'-termini of genes or operons. Interestingly, an increasing number of examples are being observed for intrinsic terminators located within 5'-UTRs.

Amongst the wide variety of genetic regulatory strategies employed by bacteria there is a growing class of examples wherein RNA polymerase responds to a termination signal within the 5'-UTR in a regulated fashion (T. M. Henkin, *Current Opinion in Microbiology* 2000, 3, 149). During certain conditions the RNA polymerase complex is directed by external signals either to perceive or to ignore the termination signal. Although transcription initiation might occur without regulation, control over mRNA synthesis (and of gene expression) is ultimately dictated by regulation of the intrinsic terminator. Generally, one of at least two mutually exclusive mRNA conformations results in the formation or disruption of the RNA structure that signals transcription

termination. A trans-acting factor, which in some instances is a RNA (F. J. Grundy, et al., *Proceedings of the National Academy of Sciences of the United States of America* 2002, 99, 11121; T. M. Henkin, C. Yanofsky, *Bioessays* 2002, 24, 700) and in others is a protein (J. Stulke, *Archives of Microbiology* 2002, 177, 433), is generally required for receiving a particular intracellular signal and subsequently stabilizing one of the RNA conformations. Riboswitches offer a direct link between RNA structure modulation and the metabolite signals that are interpreted by the genetic control machinery. A brief overview of the FMN riboswitch from a *B. subtilis* mRNA is provided below to illustrate this mechanism.

10 **i. A natural aptamer for FMN**

A highly conserved RNA domain, referred to as the *RFN* element, was identified in bacterial genes involved in the biosynthesis and transport of riboflavin and FMN (M. S. Gelfand, et al., *Trends in Genetics* 1999, 15, 439; A. G. Vitreschak, et al., *Nucleic Acids Research* 2002, 30, 3141). This element is required for genetic manipulation of the *ribDEAHT* operon (hereafter, '*ribD*') of *B. subtilis*, as mutations resulted in a loss of FMN-mediated regulation (Y. V. Kil, et al., *Molecular & General Genetics* 1992, 233, 483; V. N. Mironov, et al., *Molecular & General Genetics* 1994, 242, 201). These data led to the proposal that either a protein-based FMN sensor, or FMN itself (G. D. Stormo, Y. Ji, *Proceedings of the National Academy of Sciences of the United States of America* 2001, 98, 9465) interacts with the *RFN* element in order to repress *ribD* gene expression. However, there was no understanding of how such interactions would take place or the mechanism by which expression would be affected. Although RNA sequences that specifically bind FMN had been identified through directed evolution experimentation (C. T. Lauhon, J. W. Szostak, *Journal of the American Chemical Society* 1995, 117, 1246, M. Roychowdhury-Saha, et al., *Biochemistry* 2002, 41, 2492), they exhibit no obvious resemblances to the *RFN* element.

25 **a. Structural probing reveals FMN-mediated RNA structure modulation**

Each internucleotide linkage in a RNA polymer is susceptible to spontaneous hydrolysis by an S_N2 -like mechanism, wherein the 2' oxygen attacks the adjacent phosphorus center, leading to chain cleavage. This reaction requires a 180° orientation between the attacking nucleophile, the phosphorus center, and the 5'-oxygen leaving group (in-line conformation) (G. A. Soukup, R. R. Breaker, *RNA* 1999, 5, 1308; V.

Tereshko, et al., *RNA* 2001, 7, 405). Nucleotides that are base-paired, or otherwise structurally constrained, are typically incapable of adopting this configuration and therefore display low rates of spontaneous cleavage. In contrast, nucleotides that are structurally unrestrained exhibit much higher rates of spontaneous cleavage. These observations have been exploited in a structural probing method, referred to as “in-line probing”, which establishes the relative rates of spontaneous cleavage for a given RNA polymer and correlates this with secondary- and tertiary-structure models (V. Tereshko, et al., *RNA* 2001, 7, 405).

To assess whether the RFN element of *ribD* was responsive to FMN, a fragment of the corresponding 5'-UTR was 5'-³²P labeled and incubated in the absence and presence of FMN, and the resulting fragments were analyzed by polyacrylamide gel electrophoresis (PAGE). Interestingly, patterns differ between reactions with and without FMN, signifying that there is a structural rearrangement of the RNA upon FMN binding to *ribD*. The spontaneous cleavages of certain nucleotide positions located within inter-helical regions of the RFN element become significantly reduced in the presence of FMN, suggesting that these nucleotides are involved in forming an FMN-RNA complex, which forces structural constraints upon the RNA (Figure 12). It is this type of structural modulation that can be harnessed by the expression platform for allosteric modulation of gene expression.

Additional evidence for direct binding of FMN by the *ribD* RFN element was generated by enzymatic probing. Oligonucleotides predicted to anneal with the RFN element were added to *ribD* transcripts in the presence and absence of FMN, and the resulting mixtures were digested with RNase H (which specifically cleaves RNA:DNA heteroduplexes) and analyzed by PAGE (A. S. Mironov, et al., *Cell* 2002, 111, 747). A significant portion of transcripts bind certain oligonucleotides in the absence of FMN, but not in the presence of FMN, indicating that FMN stabilizes a structural rearrangement of *ribD* transcripts that in turn prevents annealing of the oligonucleotide.

b. Affinity and specificity of the FMN-*ribD* complex

If the RFN element serves as an aptamer for FMN, it should exhibit characteristics of a saturable receptor that has some ability to discriminate against related ligands. To obtain values for apparent dissociation constant (apparent K_D) for FMN, in-line probing assays were repeated with trace amounts of *ribD* RNA and increasing concentrations of FMN; the ligand concentration that correlates with half-maximal

modulation of RNA structure should reflect the apparent K_D . These experiments indicate that the *ribD* RNA contains a saturable ligand-binding site that exhibits an apparent K_D of ~5 nM. Furthermore, the RNA discriminates against the dephosphorylated form of FMN (riboflavin) by approximately three orders of magnitude. This exceptional ligand specificity of the *ribD* mRNA is surprising since the aptamer must generate a binding pocket for FMN that makes productive interactions with a phosphate group.

ii. FMN-induced transcription termination

a. *In vitro* transcription termination mediated by an FMN riboswitch

The relative amounts of the major transcription products for the *ribD* leader region were examined by *in vitro* transcription using T7 RNA polymerase or *Bacillus subtilis* RNA polymerase. The *ribD* leader region contains a classical intrinsic terminator just upstream of the *ribD* coding region. Interestingly, transcripts that terminated at the intrinsic terminator are specifically induced by FMN, in the absence of additional protein factors. Furthermore, mutations in the *RFN* element abrogate this phenomenon. The left-half of the terminator sequence forms alternative base-pairing interactions with a portion of the *RFN* element, thereby forming an antiterminator element. Sequence alterations of the intrinsic terminator eliminate FMN-induced termination while alterations in the antiterminator result in constitutive termination. Taken together, these observations are consistent with a mechanistic model wherein FMN directly interacts with *ribD* transcripts during conditions of excess FMN. Complex formation subsequently induces transcription termination within the 5'-UTR (Figure 12), which precludes gene expression by preventing the ORF from being transcribed. During conditions of limiting FMN, an antiterminator structure is formed within the *ribD* nascent transcript, which allows for synthesis of the downstream genes.

b. FMN-mediated control of transcription termination in vivo

The molecular details of riboswitch-mediated transcription termination are likely to be more complex than this rather simplistic model implies. For example, given that the 'decision' to form the terminator or antiterminator conformation occurs only once during transcription, the regulatory mechanism is likely to rely on precise transcriptional kinetics as well as the appropriate RNA folding pathways. Moreover, the kinetics of FMN interacting with the RNA receptor is likely a critical factor. Although the affinity that the RNA has for FMN is exceptionally strong compared to engineered aptamers, it is possible that the kinetics of ligand association might be the more important determinant

of genetic regulation. Indeed, all of these parameters are likely to conspire together in order to exert appropriate control over the intrinsic terminator. In adapting and designing riboswitches for use as described herein, the impact of transcription speed should be taken into account.

5 **iii. Control of transcription termination by other riboswitches**

Intrinsic terminators can be identified via computer-assisted search algorithms (F. Lillo, et al., 2002, 18, 971). Using such bioinformatic analyses, a subset of riboswitch RNAs that are predicted to contain an intrinsic terminator and an alternate antiterminator structural element can be identified (M. Mandal, et al., *Cell* 2003, 113; A. G. Vitreschak, et al., *Nucleic Acids Research* 2002, 30, 3141; F. J. Grundy, T. M. Henkin, *Molecular Microbiology* 1998, 30, 737; S. Kochhar, H. Paulus, *Microbiology* 1996, 142, 1635; D. A. Rodionov, et al., *Journal of Biological Chemistry* 2002, 277, 48949). Therefore, the results described above for the FMN riboswitch are indicative of the mechanisms used by many other riboswitch RNAs. Indeed, SAM- and TPP-dependent riboswitches have been demonstrated to exert control over termination via formation of mutually exclusive intrinsic terminator and antiterminator structures (see, e.g., Example 7). Furthermore, mutations that disrupt and subsequently restore helices within the SAM riboswitch aptamer result in loss and restoration, respectively, of SAM binding. Concurrently, these mutations also result in disruption or restoration of SAM-induced transcription termination in accordance with ligand-binding function. Riboswitches can be adapted and designed to exert control over transcription termination signals that differ appreciably from classical intrinsic terminators according to principles described herein. As described elsewhere herein, expression platform domains having expression-controlling stem structures can be matched to aptamer domains by designing the P1 stem of the aptamer domain such that the control strand (P1b) of the aptamer can form a stem structure with the regulated strand (P1c) of the expression platform.

20 **4. Riboswitch Regulation of Translation Initiation in Bacteria**

An alternative mechanism of genetic control by riboswitches is the modulation of translation initiation. Unlike transcription termination, the entire mRNA would be synthesized by RNA polymerase, but expression would be prevented by the riboswitch until the metabolite concentration reached a certain level. In most instances, it was observed that riboswitches prevent translation initiation in the presence of high concentrations of target metabolite. However, riboswitches can be designed and adapted

such that allosteric modulation of riboswitch structures could lead to translation activation. The regulatory mechanism of translation control is briefly described below for a TPP riboswitch from *E. coli*.

i. A natural aptamer for TPP

5 A conserved RNA element, referred to as the *thi* box, was identified within 5'-UTRs of mRNAs that are responsible for thiamine biosynthesis and transport (D. A. Rodionov, et al., *Journal of Biological Chemistry* 2002, 277, 48949; J. Miranda-Rios, M. Navarro, M. Soberon, *Proceedings of the National Academy of Sciences of the United States of America* 2001, 98, 9736.). Genetic experiments confirmed that this structural
10 element was required for thiamine-dependent regulation of *Rhizobium meliloti* thiamine biosynthesis genes (J. Miranda-Rios, M. Navarro, M. Soberon, *Proceedings of the National Academy of Sciences of the United States of America* 2001, 98, 9736), yet no regulatory factor had been identified through classical genetic experimentation. Therefore, it was possible that the *thi* box might serve as a portion of a riboswitch that
15 responds to thiamine or its derivatives.

In *E. coli*, thiamine biosynthesis and transport genes are primarily located within three operons and four single genes (T. P. Begley, et al., *Archives of Microbiology* 1999, 171, 293), wherein each operon is preceded by a *thi* element. To begin to assess the regulatory properties of these sequences, the leader regions for the *thiMD* and
20 *thiCEFSGH* operons were utilized to construct transcriptional and translational fusions to a *lacZ* reporter gene (see Example 2). Addition of exogenous thiamine results in repression of the *lacZ* reporter gene in *E. coli*. Results from these data demonstrate that the *thiM* gene is regulated primarily at the level of translation while the *thiC* leader region confers both transcriptional and translational regulation to the *lacZ* reporter.

a. Direct binding of thiamine pyrophosphate by *E. coli* mRNAs

As described above for the FMN aptamer, direct binding of TPP to the *thiM* and
25 *thiC* leaders was demonstrated by in-line probing assays (see Example 2). The addition of thiamine, thiamine monophosphate (TP), or the pyrophosphate derivative (TPP) leads to structural rearrangement of the *thiM* RNA, particularly in the region encompassing the
30 *thi* element (Figure 13). Significantly, TPP, which is the bioactive form of thiamine, exhibits the best affinity between the ligands, with an apparent K_D of 500 nM, while TP and thiamine associate to *thiM* with apparent K_D values of 3 μ M and 40 μ M,

respectively. In-line probing assays of RNAs resembling the *thiC* leader region reveal even more dramatic discrimination between thiamine and its phosphorylated forms, exhibiting greater than a 1,000-fold difference between binding of thiamine and TPP. These data are consistent with genetic experiments that suggested that TPP synthesis was required for regulation (E. Webb, et al., *Journal of Bacteriology* 1996, 178, 2533; E. Webb, D. Downs, *Journal of Biological Chemistry* 1997, 272, 15702). Also, this system provides another example of a natural RNA aptamer that makes productive contacts to phosphate groups.

b. Confirmation of TPP binding by equilibrium dialysis

10 RNAs resembling the *thiM* leader region were synthesized and placed into one side of a two-chamber equilibrium dialysis apparatus, in which the compartments are separated by a 3000-dalton molecular-weight-cut-off dialysis membrane. ³H-thiamine was preferentially retained within the *thiM*-containing chamber when allowed to equilibrate between chambers (see Example 2). This effect could be eliminated by
15 providing excess unlabeled thiamine, but could not be reversed when supplemented with oxythiamine, a close chemical analog of thiamine. Additionally, a mutated version of *thiM* was unable to shift ³H-thiamine to the RNA-containing chamber. Together, these data are indicative of the formation of stable *thiM*:thiamine complexes, wherein the sequence of the RNA and the chemical form of the ligand are critical for maximal
20 binding affinity.

ii. Binding of thiamine derivatives correlates with structural modulation

Close inspection of in-line probing data for *thiM* reveal two surprising patterns of structural modulation. First, the relative rates of spontaneous fragmentation between reactions containing either thiamine or TPP differ within an internal loop of the *thi*
25 element (Figure 13). Nucleotides in this region adopt an increase in structural order in the presence of TPP but not with thiamine, implying this region is somehow involved in formation of a pyrophosphate-recognition pocket. Secondly, the region of the SD sequence is the only portion outside of the *thi* element that becomes structurally modulated in the presence of TPP.

30 Specifically, the SD sequence exhibits a significant decrease in spontaneous cleavage relative to reactions lacking TPP, suggesting that the SD is converted into a more structurally constrained form upon binding of TPP. This idea is consistent with a mechanism (Figure 13) whereby in the absence of TPP the SD has a significant degree of

single-stranded character and is accessible for translation initiation. An anti-SD sequence is proposed to interact with an anti-anti-SD sequence within the TPP aptamer under these conditions. In contrast, during conditions of excess TPP, a TPP-RNA complex is formed that disrupts the base pairing of the anti-SD sequence, which is then free to interact
5 directly with the SD and decrease the single-stranded character of the region, hence decreasing efficiency of translation initiation. Preliminary site-directed mutagenesis of the *thiM* mRNA supports this overall model (see Example 2). Specifically, mutations that disrupt TPP binding also disrupt regulation of translation for *thiM-lacZ* fusions, while mutations that alter the anti-SD sequence affect regulation but do not affect TPP
10 binding. Thus, binding of thiamine correlates with both the structural accessibility of the SD and the translation efficiency *in vivo*.

iii. Control of translation initiation by other riboswitches

Bioinformatics analyses are consistent with molecular mechanisms similar to that of *thiM* also being recurrent amongst riboswitch RNAs. Specifically, anti-SD and anti-
15 anti-SD structures have been proposed for several riboswitch classes, including FMN (A. G. Vitreschak, et al., *Nucleic Acids Research* 2002, 30, 3141), lysine, TPP (D. A. Rodionov, et al., *Journal of Biological Chemistry* 2002, 277, 48949), coenzyme B₁₂ (see Example 1) and SAM. In general, riboswitches from Gram-negative organisms seem to favor expression platforms that exert control over translation, while riboswitches from
20 Gram-positive bacteria appear to predominately use expression platforms that control transcription termination. The latter can reflect a greater reliance upon multigene transcriptional units in Gram-positive organisms, which might be more efficient to preclude transcription of long operons when the gene products are unnecessary.

Biochemical evidence for riboswitch-mediated control over translation initiation
25 has also been obtained for FMN and AdoCbl riboswitches (see Example 1). FMN binding to a riboswitch that regulates the *B. subtilis ypaA* gene results in alteration of the SD structural context, similar to what was observed for *thiM*. Interestingly, this genetic control element has also been proposed to regulate *ypaA* transcription (J. M. Lee, et al., *Journal of Bacteriology* 2001, 183, 7371), although the leader region does not contain an
30 obvious intrinsic terminator structure. Binding of AdoCbl to the *E. coli btuB* riboswitch has also been demonstrated to correlate with regulation of translation *in vivo*.

Certain riboswitch RNAs exert control over transcription and translation using the same RNA sequence. For this class of riboswitches, the SD sequence is contained

within an intrinsic terminator. Therefore, the formation of the terminator structure also enacts formation of a SD-sequestering structure. In total, all of these observations suggest that although the *thiM* and *ribD* riboswitches represent useful paradigms for riboswitch-mediated control of translation and transcription, respectively, there are likely to be a wide variety of molecular mechanisms utilized by riboswitch RNAs for control of gene expression. Indeed, TPP riboswitches that must be employing different mechanisms of control have been identified in several plant and fungal species (see Example 4). The placement of these RNAs near splice sites in some instances and in the 3'-UTR in others indicate TPP-responsive control over splicing and mRNA stability or expression, respectively.

5. Early origins?

The FMN, TPP, lysine and AdoCbl riboswitch RNAs are widespread among evolutionarily distant microorganisms, implying an ancient origin for these RNA genetic elements (A. G. Vitreschak, et al., *Nucleic Acids Research* 2002, 30, 3141; D. A. Rodionov, et al., *Journal of Biological Chemistry* 2002, 277, 48949; D. A. Rodionov, et al., *Journal of Biological Chemistry* 2002, 277, 48949). SAM, guanine, and adenine riboswitches are also represented in numerous different genera, although they appear to be primarily limited to Gram-positive bacteria, with a few Gram-negative bacteria as exceptions (see Example 6). In all instances, the structural and sequence conservation of riboswitch classes is limited to the aptamer domain (Figure 11). This is not unexpected given that the aptamer RNA must preserve its capability to bind the target chemical, which has not been significantly modified through evolution. In contrast, there is considerable sequence and structural diversity between expression platforms, even between riboswitches of the same class and within the same organism. Together, these data hint that the ligand-binding properties of riboswitch aptamer domains have been maintained throughout expansive evolutionary timescales.

Furthermore, the ligands for riboswitch RNAs have been proposed to be functional relics from a hypothetical RNA-based world, in which RNA polymers provided all the necessary catalytic and genomic content for some of the earliest self-replicating organisms (H. B. White, 3rd, *Journal of Molecular Evolution* 1976, 7, 101; G. F. Joyce, *Nature* 2002, 418, 214). Therefore it is tempting to speculate that as cofactor-binding RNAs the aptamer domains from riboswitches may have been useful in the context of an RNA-based world for some of the earliest forms of genetic control, for

allosteric modulation of ribozymes, or as part of ribozymes that utilized the ligands as catalytic cofactors.

6. Riboswitches as drug targets and genetic tools

Riboswitches are utilized for control of numerous genes involved in the biosynthesis and transport of prokaryotic enzymatic cofactors. At least 69 genes, which represents nearly 2% of *Bacillus subtilis* total genomic content, is under control of riboswitch RNAs (Table 1), exemplifying the extensive use of riboswitch RNAs for genetic control in prokaryotes. (M. Mandal, et al., *Cell* 2003, 113). Many riboswitch-mediated genes are expected to be essential under most growth conditions. Interference with riboswitch function is then predicted to result in dramatic destabilization of vital metabolic pathways and perhaps, cessation of growth. Therefore, it seems likely that compounds that closely resemble the target metabolites will bind to riboswitch RNAs and cause a decrease in gene expression. If this analog-induced disruption of gene expression is sufficient, then such compounds might be candidates for antimicrobial applications.

Table 1

Ligand	Transcriptional Unit	Predicted Gene Function(s)
Lysine	<i>lysC</i>	Aspartokinase II
Flavin mononucleotide	<i>ypaA</i>	Putative flavin transporter
	<i>ribD-ribE-ribBA-ribH</i>	Riboflavin biosynthesis
Adenosylcobalamin	<i>yvrC-yvrB-yvrA-yvqK</i>	Unknown; similar to iron transport proteins
Thiamine pyrophosphate	<i>thiC</i>	Biosynthesis of thiamine pyrimidine moiety
	<i>tenA1-thiX1-thiY1-thiz1-thiE2-thiO-thiS-thiG-thiF-thiD</i>	Thiamine biosynthesis
	<i>ykoF-ykoE-ykoD-ykoC</i>	Unknown
	<i>yuaJ</i>	Unknown; putative thiamine transporter
	<i>ylmB</i>	Similar to acetylornithine deacetylase
Guanine	<i>yxjA</i>	Similar to pyrimidine nucleoside transport
	<i>xpt-pbuX</i>	Xanthine permease

	<i>pbuG</i>	Hypoxanthine/Guanine permease
	<i>purE-purK-purB-purC-purS-purQ-purL-purF-purM-purN-purH-purD</i>	Purine biosynthesis
Adenine	<i>ydhL</i>	Unknown
<i>S</i> -adenosylmethionine	<i>yitJ</i>	Putative methylene tetrahydrofolate reductase
	<i>metI-metC</i>	Methionine biosynthesis
	<i>ykrT-ykrS</i>	5' methylthioadenosine recycling pathway
	<i>ykrW-ykrX-ykrY-ykrZ</i>	5' methylthioadenosine recycling pathway
	<i>cysH-cysP-sat-cysC-ylnD-ylnE-ylnF</i>	Cysteine biosynthesis
	<i>yoaD-yoaC-yoaB</i>	Unknown
	<i>metE</i>	Methionine synthase, B ₁₂ -independent
	<i>metK</i>	<i>S</i> -adenosylmethionine synthetase
	<i>yusC-yusB-yusA</i>	Unknown ABC transporter
	<i>yxjG</i>	Unknown
	<i>yxjH</i>	Unknown

Table 1. Distribution of known riboswitch classes in *Bacillus subtilis*. Gene nomenclature is derived from the SubtiList database except for *metI* and *metC*, which are recent designations (S. Auger, et al., *Microbiology* 2002, 148, 507). Functional roles for *ypaA* (R. A. Kreneva, et al., *Genetika* 2000, 36, 1166), *yuaJ* (D. A. Rodionov, et al., *Journal of Biological Chemistry* 2002, 277, 48949), *ykrTS* (B. A. Murphy, et al., *Journal of Bacteriology* 2002, 184, 2314), and *ykrWXYZ* (B. A. Murphy, et al., *Journal of Bacteriology* 2002, 184, 2314.), have recently been proposed.

There is clear precedence for the targeting of RNAs with small molecule drugs (G. J. Zaman, et al., *Nucleic Acids Research* 2002, 30, 62), the most obvious example being that of ribosomal RNA. Several other bacterial-specific RNAs have been explored as candidates for small molecule drug interaction; however, the approach relies upon screening large chemical libraries for those chemicals that fortuitously interact with the RNA of interest, even though the RNA itself does not naturally form a binding pocket for small organic molecules. Riboswitch RNAs therefore exhibit an advantage in

antimicrobial development given that they serve as a receptor for small molecule ligands, much like their protein receptor counterparts.

In addition to their use as targets for chemical inhibition, understanding of the mechanisms utilized by natural riboswitch RNAs allows adaptation of riboswitches and development of new riboswitches as novel genetic control elements. Numerous aptamer RNA sequences have been identified that interact with a wide variety of small organic molecules (M. Famulok, *Current Opinion in Structural Biology* 1999, 9, 324). Engineered riboswitches can be generated that respond to non-biological, or otherwise metabolically inert, compounds. Such genetic control elements can be used for a variety of expression control and molecular detection applications.

D. Example 4: Eukaryotic Riboswitches

1. Abstract

Genetic control by metabolite-binding mRNAs is wide spread in prokaryotes. These "riboswitches" are typically located in non-coding regions of mRNA, where they selectively bind their target compound and subsequently modulate gene expression. Disclosed are mRNA elements that have been identified in fungi and in plants that match the consensus sequence and structure of thiamine pyrophosphate-binding domains of prokaryotes. In *Arabidopsis*, the consensus motif resides in the 3'-UTR of a thiamine biosynthetic gene, and the isolated RNA domain binds the corresponding coenzyme in vitro. These results suggest that metabolite-binding mRNAs possibly are involved in eukaryotic gene regulation and that some riboswitches might be representatives of an ancient form of genetic control.

2. Introduction

Riboswitches are genetic control elements that can be found in the 5'-untranslated region of certain messenger RNAs of prokaryotes (see Examples 1-3). These genetic switches exhibit two surprising properties. First, the mRNA is able to form a highly selective binding site for the target metabolite without the aid of proteins. Second, metabolite binding brings about an allosteric reorganization of RNA structure that leads to alterations in genetic expression. Unlike many other genetic control systems, riboswitches do not require metabolite-binding proteins to serve as sensors, and thus offer a direct link between the genetic information that is encoded by an mRNA and its chemical surroundings.

A number of distinct types of riboswitches have been confirmed by biochemical and genetic analyses. For example, a coenzyme B₁₂-binding RNA has been shown (Example 1) to control expression of the *Escherichia coli* *btuB* gene, which encodes a cobalamin transport protein. Riboswitches triggered by thiamine pyrophosphate (TPP) have been shown to control operons in *E. coli* (Example 3) and *Bacillus subtilis* (Example 6) that are responsible for biosynthesis of this coenzyme. In addition, the *RFN* element, which frequently is found in the 5'-untranslated region of genes responsible for the biosynthesis or import of riboflavin and FMN, serves as the receptor portion of FMN-dependent riboswitches in *Bacillus subtilis* (see Examples 3 and 6). Recently, it has been determined that certain S-box motifs that are located in the 5'-UTRs of numerous genes in *B. subtilis* bind the coenzyme S-adenosylmethionine (SAM) with high affinity and precision. These findings indicate that riboswitches are used to recognize a diverse collection of metabolites and that direct sensing of small molecules by mRNAs is an important form of genetic control for certain organisms. Disclosed herein, is evidence that metabolite-binding domains are embedded in certain mRNAs of eukaryotes, indicating that higher organisms might also exploit riboswitches for genetic control.

3. Results

Disclosed are many RNA elements that have been identified in prokaryotes that exhibit sequence similarity to the B₁₂- and SAM-dependent riboswitches. Given the relatively large size and sequence complexity of these RNA motifs, it is unlikely that numerous evolutionary reinventions of the same elements would have occurred. Furthermore, the metabolite triggers of these genetic switches are predicted to have been present in a time before the emergence of proteins (White, 1976; Benner et al., 1989; Jeffares et al., 1998). This is consistent with the known classes of metabolite-sensing RNAs having originated in the ancient RNA world, which is believed to be a time before the emergence of proteins and when metabolism was guided entirely by RNA (Joyce, 2002).

If the present-day riboswitches are of ancient origin, then eukaryotes might possess RNA genetic switches that are descendent from the last common ancestor of modern cells. Disclosed herein several eukaryotes carry RNA domains that conform to the consensus sequence and structure of the metabolite-binding domain of the TPP riboswitch class (Fig. 14A) (The mRNAs that carry the TPP-binding domains encode for

a protein that is homologous to the thiC protein of *E. coli*. This protein enzyme catalyzes the conversion of 5-aminoimidazole ribotide (AIR) to hydroxymethyl pyrimidine phosphate (HMP-P), which is a key biosynthetic step in the synthesis of thiamine and ultimately TPP (Vander Horn et al., 1993; Begley et al., 1999)). For example, a putative
5 thiamine biosynthesis gene of *Arabidopsis thaliana* carries an RNA element (Fig. 14B) in its 3'-UTR that conforms to the consensus TPP-binding domain. Similar RNA elements are found in rice (*Oriza sativa*) and bluegrass (*Poa secunda*). RNA elements that conform to the TPP-binding sequence and structure are also present in fungi such as *Neurospora crassa* (Fig. 14C) and *Fusarium oxysporum*. As with plants, the riboswitch
10 homologs in fungi are located in genes that have been implicated in the biosynthesis of thiamine, suggesting that in each case their role is to maintain required coenzyme levels by modulating expression of the appropriate biosynthetic genes. A sequence alignment of the homologous domains found in eukaryotes compared to that of the gram negative bacterium *E. coli* (*thiC* and *thiM*) and the gram positive bacterium *Chlostridium*
15 *acetobutylicum* (*thiC*) is depicted in Fig. 15.

The RNA element corresponding to the consensus TPP-binding domain of *A. thaliana* (Fig. 14A) was generated by in vitro transcription of a synthetic DNA template and the RNA was subjected to "in-line probing" (Fig. 16A). This method relies on the spontaneous breakdown of RNA phosphodiester linkages, whose pattern of cleavage can
20 be used to reveal the structural and functional features of ligand-binding RNAs (see Examples 1-3). Indeed, the riboswitch-like element exhibits TPP-dependent structural modulation and has a fragmentation pattern that is consistent with the predicted secondary structure of TPP riboswitches from bacteria (see Examples 2 and 3). In addition, this structure-probing method has been used herein to establish that the RNA
25 binds TPP with an apparent dissociation constant (K_D) of ~50 nM (Fig. 16B), which is similar to that determined previously for an *E. coli* riboswitch variant. Similarly, it has been demonstrated that the sequence elements of fungi that correspond to the TPP riboswitch consensus also bind TPP with high affinity.

Sequestering of the ribosome binding site and transcription termination are
30 demonstrated mechanisms for TPP riboswitches in *E. coli* (Fig. 17). Since the TPP-binding element in plants is located immediately upstream from the polyA tail, it is possible that metabolite binding might regulate mRNA processing and stability. Alternatively, a consensus TPP-binding sequence (Fig. 14C) identified in the fungal

genome of *N. crassa* resides in an intron, suggesting that RNA splicing might also be guided by metabolite-binding pre-mRNAs. In prokaryotes, ligand binding typically brings about allosteric changes in the Watson-Crick base pairing arrangements near gene control elements such as transcription terminators and ribosome binding sites. Likewise, secondary structure rearrangements by metabolite-binding riboswitches can be used to modulate a greater variety of RNA processing, transport and expression pathways in eukaryotes.

Although it is likely that TPP-binding domains and those for coenzyme B₁₂, FMN, and SAM are of ancient origin, it is possible that other examples of metabolite-binding mRNAs have emerged more recently in evolution. These newer riboswitches would be more narrowly distributed across the phylogenetic landscape, so efforts to search for new riboswitches that are triggered by compounds that are not ancient and universally distributed will be difficult. Regardless of the scope of riboswitch use in modern organisms, both natural and engineered riboswitches could have significant utility. Given the central role that known riboswitches serve in modulating the concentration of key coenzymes, these RNAs can serve as new targets for drug discovery efforts. Therefore, reverse engineering of natural riboswitches can be used to establish a conceptual basis for creating designer riboswitches for the purposeful control of eukaryotic genes.

E. Example 5: Lysine Riboswitches

The precise control of gene expression in response to changes in the chemical and physical environment of cells requires selective interactions between biochemical sensor elements and the molecules that carry or interpret genetic information. Most known genetic factors that respond to such environmental changes are proteins (Ptashne and Gann 2002). However, a number of studies (e.g. see Examples 1-3 and 6-8) have demonstrated that natural RNA molecules can also recognize small organic compounds and harness allosteric changes to control the expression of adjacent genes. These metabolite-binding RNA domains, termed riboswitches, typically are embedded within the 5'-UTRs of mRNAs and control the expression of proteins involved in the biosynthesis or import of the target compound. Riboswitches also play an important role in controlling fundamental metabolic pathways in bacteria involved in sulfur metabolism, and in the biosynthesis of various coenzymes and purines (see Example 6). Furthermore, riboswitches are phylogenetically widespread amongst eubacterial

organisms, and both sequence and biochemical data suggest that riboswitches are also present in the genes of eukaryotes (see Example 4).

These observations indicate that riboswitches likely comprise a widely used mechanism of genetic control in living systems. Transcription of the *lysC* gene of *B. subtilis* is repressed by high concentrations of lysine (Kochhar, S., and Paulus, H. 1996, *Microbiol.* 142:1635-1639; Mäder, U., et al., 2002, *J. Bacteriol.* 184:4288-4295; Patte, J.C. 1996. Biosynthesis of lysine and threonine. In: *Escherichia coli and Salmonella: Cellular and Molecular Biology*, F.C. Neidhardt, et al., eds., Vol. 1, pp. 528-541. ASM Press, Washington, DC; Patte, J.-C., et al., 1998, *FEMS Microbiol. Lett.* 169:165-170), but that no protein factor had been identified that served as the genetic regulator (Liao, H.-H., and Hseu, T.-H. 1998, *FEMS Microbiol. Lett.* 168:31-36). The *lysC* gene encodes aspartokinase II, which catalyzes the first step in the metabolic pathway that converts L-aspartic acid into L-lysine (Belitsky, B.R. 2002. Biosynthesis of amino acids of the glutamate and aspartate families, alanine, and polyamines. In: *Bacillus subtilis and its Closest Relatives: from Genes to Cells*. A.L. Sonenshein, J.A. Hoch, and R. Losick, eds., ASM Press, Washington, D.C.). Interestingly, several efforts have been successful in generating mutants that exhibit constitutive expression of the aspartokinase II enzyme, and all mutations map to the 5'-UTR of the *lysC* mRNA (Boy, E., et al., 1979. *Biochimie* 61:1151-1160; Lu, Y., et al., 1991, *J. Gen. Microbiol.* 137:1135-1141; Lu, Y., et al., 1992, *FEMS Microbiol. Lett.* 92:23-27). Furthermore, a significant level of sequence similarity was identified between the *B. subtilis* and *E. coli lysC* 5'-UTRs (Patte, J.-C., et al., 1998, *FEMS Microbiol. Lett.* 169:165-170.). These characteristics are consistent with a lysine-responsive riboswitch serving as the genetic control element for this gene.

1. Materials and methods

i. Chemicals and Oligonucleotides

L-lysine, all analogs with the exception of L- α -homolysine (compound 6, Fig. 20A), tritiated lysine (L-Lysine-[4,5-³H(N)]), and the four dipeptides were purchased from Sigma. A protocol adapted from that reported previously (Dong, Z. 1992, *Tetrahedron Lett.* 33:7725-7726) was used to synthesize L- α -homolysine. Purity and integrity of synthetic L- α -homolysine was confirmed by TLC and NMR.

DNA oligonucleotides were synthesized by the HHMI Keck Foundation Biotechnology Resource Center at Yale University, purified by denaturing PAGE and eluted from the gel by crush-soaking in 10 mM Tris-HCl (pH 7.5 at 23°C), 200 mM NaCl, and 1 mM EDTA. Oligonucleotides were recovered from solution by precipitation with ethanol.

ii. Phylogenetic Analyses

L box domains were identified by sequence similarity to the *B. subtilis* *lysC* 5'-UTR. Ultimately, the program was used to search for degenerate matches to the pattern (WAGAGGNGC [10] A [3] RKTA [50] RRGR [10] CCGARR [40] GG [13] VAA [13] YTGTC A [36] TGRWG [2] CTWY), however, less complete versions of this pattern were used with iterative refinements to identify the consensus sequence and structure of the L box motif. Bracketed numbers are variable gaps with constrained maximum lengths denoted. Nucleotide notations are as follows: Y = pyrimidine; R = purine; W = A or T; K = G or T; V = A, G or C. Up to six violations of this pattern were permitted when forming the phylogeny depicted in Figure 18.

iii. In-line Probing of RNA Constructs

The *B. subtilis* 315 *lysC*, 237 *lysC* and 179 *lysC* RNAs were prepared by *in vitro* transcription using T7 RNA polymerase and the appropriate PCR DNA templates. RNA transcripts were dephosphorylated and subsequently 5' ³²P-labeled using a protocol similar to that described previously (Seetharaman, S. et al., 2001, *Nature Biotechnol.* 19, 336-341). Labeled precursor RNAs (~2 nM) were subjected to in-line probing using conditions similar to those described in Examples 1 and 2. Reactions (10 µL) were incubated for 40 hr at 25°C in a buffer containing 50 mM Tris (pH 8.5 at 25°C), 20 mM MgCl₂ and 100 mM KCl in the presence or absence of L-lysine or various analogs as indicated for each experiment. Denaturing 10% PAGE was used to separate spontaneous cleavage products, which were detected and quantitated by using a Molecular Dynamics PhosphorImager and ImageQuaNT software.

iv. Equilibrium Dialysis and Scatchard Analyses

Equilibrium dialysis assays were conducted using a DispoEquilibrium Dialyzer (ED-1, Harvard Bioscience), wherein two chambers *a* and *b* were separated by a 5,000 MWCO membrane. The final composition of buffer included 50 mM Tris-HCl (pH 8.5 at 25°C), 20 mM MgCl₂ and 100 mM KCl (30 µL delivered to each chamber). Assays were

initiated by the addition of ^3H -lysine (50 nM initial concentration prior to equilibration; 40 Ci mmol $^{-1}$; 15,000 cpm) to chamber *a*. When present, RNA (179 *lysC*) was introduced into chamber *b* to yield a concentration of 10 μM . After 10 hr of equilibration at 25°C, a 3- μl aliquot from each chamber was removed for quantitation by liquid scintillation
5 counter. Competition assays were established by delivering an additional 3 μL of buffer to *a* and an equivalent volume of buffer containing 50 μM unlabeled L-lysine, D-lysine, L-oronithine, or L-lysine hydroxamate as indicated to *b*. After 10 hr of additional incubation at 25°C, 3- μl aliquots were again drawn for quantitation of tritium distribution.

10 Scatchard data points were generated as described above with the following exceptions. RNA was added to chamber *b* to yield a concentration of 1 μM RNA and equilibration of the dialysis mixtures proceeded for 20 hr. In addition, ^3H -lysine concentrations were varied from 50 nM to 2.5 μM . Calculation of points on the Scatchard plot from the equilibrium dialysis data was carried out as described elsewhere
15 herein.

v. In vitro Transcription Termination Assays

Transcription termination assays were conducted using a method of single-round transcription adapted from that described previously (Landick, R., et al., 1996, *Methods Enzymol.* 274:334-353). The template for *lysC* 5'-UTR transcription was altered (C6G of
20 the RNA) such that the first C residue of the nascent RNA is not encountered until position 17. Polymerization was initiated by the addition of a mixture of ApA dinucleotide (1.35 μM), GTP and UTP (2.5 μM each) plus unlabeled ATP (1 μM) and [α - ^{32}P]-ATP (4 μCi), which was incubated for 10 min. Halted complexes are restarted by the addition of 150 μM each of the four NTPs, and heparin (0.1 mg mL $^{-1}$) is
25 simultaneously added to prevent polymerases from initiating transcription on new templates. Transcription mixtures also contained 20 mM Tris-HCl (pH 8.0 at 23°C), 20 mM NaCl, 14 mM MgCl $_2$, 0.1 mM EDTA, 0.01 mg/mL BSA, 1% v/v glycerol, 4 pmoles DNA template, 0.045 U μL^{-1} *E. coli* RNA polymerase (Epicenter, Madison, WI), and 10 mM of L-lysine or the lysine analog as indicated for each experiment. Reactions were
30 incubated for an additional 20 min at 37°C and the products were examined by denaturing 6% PAGE followed by analysis using a PhosphorImager.

vi. In vivo Analysis of *lysC* Genetic Variants

Fusions of the *lysC* 5'-UTR with a *lacZ* reporter gene were used to assess the function of the lysine riboswitch *in vivo* using methods similar to those described elsewhere herein. Briefly, the *lysC* 5'-UTR, comprising the promoter and the first 315 nucleotides of the transcription template, was prepared as an EcoRI-BamHI fragment by PCR. Sequence variants M1 through M3, G39A, and G40A were generated by PCR amplification of the wild-type construct using primers that carried the desired mutations. The PCR products were cloned into pDG1661 immediately upstream of the *lacZ* reporter gene and the integrity of the resulting clones were confirmed by sequencing.

Transformations of pDG1661 variants into *B. subtilis* strain 1A40 (obtained from the *Bacillus* Genetic Stock Center, Columbus, OH) were performed and the correct transformants were identified by selecting for chloramphenicol resistance and screening for spectinomycin sensitivity.

Cells were grown with shaking at 37°C either in rich medium (2XYT broth or tryptose blood agar base) or defined medium (0.5% w/v glucose, 2 g L⁻¹ (NH₄)₂SO₄, 18.3 g L⁻¹ K₂HPO₄·3H₂O, 6 g L⁻¹ KH₂PO₄, 1 g L⁻¹ sodium citrate, 0.2 g L⁻¹ MgSO₄·7H₂O, 5 µM MnCl₂, and 5 µM CaCl₂. Methionine, lysine, and tryptophan were added to 50 µg mL⁻¹ for routine growth. Growth under lysine-limiting conditions was established by incubation under routine growth conditions in defined medium to an A₅₉₅ of 0.1, at which time the cells were pelleted by centrifugation, resuspended in minimal medium, split into five aliquots, and supplemented with five different media types as defined in the legend to Fig. 22C. Cultures were incubated for an additional 3 hr before performing β-galactosidase assays.

2. Results

i. The L box: a conserved mRNA element that is important for genetic control

Riboswitches are typically formed by close juxtaposition of a metabolite-binding 'aptamer' domain and an 'expression platform' that interfaces with mRNA elements necessary for gene expression. Although the RNA sequences and structural components that serve as the expression platform change significantly throughout evolution, the aptamer domain largely retains the sequence composition of its ligand-binding core along with the major secondary-structure features. This permits the use of phylogenetic

analyses to identify related RNA domains and to establish a consensus sequence and structure for a given class of riboswitches.

Beginning with the sequence homology reported to exist between the *lysC* 5'-UTRs of three bacterial species (Patte, J.-C., et al., 1998, *FEMS Microbiol. Lett.* 169:165-170), the number of representatives was expanded using an algorithm that searches for related sequences and secondary structures (e.g. see Examples 4 and 6). 31 representatives of this RNA domain, termed the "L box", in the 5'-UTRs of *lysC* homologs and other genes related to lysine biosynthesis from a number of Gram-positive and Gram-negative organisms were identified (Fig. 18). The sequence alignment reveals that the RNA forms a five-stem junction wherein major base-paired domains are interspersed with 56 highly conserved nucleotides (Fig. 19A). Furthermore, the base-paired elements P2, P2a, P2b, P3 and P4 each appear to conform to specific length restrictions, suggesting that they are integral participants in the formation of a highly structured RNA. It was also noticed that conserved sequences in the junction between stems P2 and P2a conform to a "loop E" motif, which is an RNA element that occurs frequently in other highly-structured RNAs (e.g. see Leonitis, N.B., and Westhof, E. 1998, *J. Mol. Biol.* 283:571-583).

The L box domain of the *B. subtilis lysC* mRNA resides immediately upstream from a putative transcription terminator stem (Kochhar, S., and Paulus, H. 1996, *Microbiol.* 142:1635-1639; Patte, J.-C., et al., 1998, *FEMS Microbiol. Lett.* 169:165-170). In several other riboswitches with similar arrangements (e.g. Examples 3 and 6), the 5'-UTR can be trimmed to separate the minimal aptamer domain from the adjacent expression platform. An RNA fragment (237 *lysC*, Fig. 19B), encompassing nucleotides 1 through 237 of the *lysC* 5'-UTR, was generated and examined for allosteric function. This construct, which excludes the putative transcription terminator stem, was subjected to structural analysis by in-line probing (Soukup, G.A. and Breaker, R.R. 1999, *RNA* 5:1308-1325) to determine whether the presence of lysine alters RNA structure. It was observed that 237 *lysC* exhibits a pattern of spontaneous RNA cleavage (Fig. 19C) that is consistent with the secondary structure model of the L box motif constructed from phylogenetic sequence data. Furthermore, it was found that the addition of 10 μ M L-lysine causes significant changes in the cleavage pattern at four locations along the RNA chain, indicating that allosteric modulation of the 5'-UTR fragment is occurring. In addition, the same pattern of spontaneous cleavage and amino acid-dependent structural

modulation was observed when using the 179 *lysC* RNA construct, which encompasses only the most highly-conserved portion of the L-box motif (nucleotides 27 through 205 of the *lysC* 5'-UTR).

5 A reduction of spontaneous cleavage is observed in each of the four sites of metabolite-induced structural modulation. In most instances, a reduction in spontaneous cleavage is due to the nucleotides becoming more ordered in the complex formed between RNA and its ligand (Soukup, G.A. and Breaker, R.R. 1999, *RNA* 5:1308-1325). Interestingly, these four groups of nucleotides are located at the center of the 5-stem junction of the L box secondary structure model (Fig. 19B), implying that these
10 nucleotides are directly involved in recognizing the amino acid target. Similar patterns of ligand-induced structural modulation have been observed with the aptamer domains of other riboswitches (see Examples 2, 3 and 6).

ii. The lysine aptamer exhibits high specificity for L-lysine and discriminates against closely-related analogs

15 Riboswitches, like their counterpart genetic factors made of protein, must exhibit sufficient specificity and affinity for their target metabolite in order to achieve precision genetic control. To examine the molecular recognition characteristics of the *lysC* L box domain, a series of in-line probing assays were performed using various analogs of lysine at 100 μ M. The properties of a lysine analog collection were examined, wherein
20 each compound carries minimal chemical changes relative to L-lysine (Fig. 20A). Nearly every chemical alteration to the amino acid renders the compound incapable of causing a structural modulation of the 179 *lysC* RNA (Fig. 20B). Perhaps most striking is that the RNA does not undergo structural modulation in the presence of D-lysine, which differs from L-lysine by the stereochemical configuration at a single carbon center.

25 The absence of significant structural modulation in the presence of D-lysine and of other analogs indicates that at least three points of contact are being made between the RNA and its amino acid target. Specifically, the observation that analogs 1, 3, and 4 fail to induce structural modulation is consistent with contacts being made to the amino and carboxy groups of the chain atoms, and to the amino group of the side chain,
30 respectively. Moreover, the failures of compounds 2, 5, 6, 7 and 8 to induce conformational change in the RNA indicate that the aptamer forms a highly discriminating binding pocket that can measure the length and the integrity of the alkyl side chain. This high level of molecular discrimination is of particular biological

significance, as a genetic switch for lysine most likely must respond exclusively to L-lysine and not closely related natural compounds.

Similarly, the allosteric response of the 179 *lysC* RNA to various dipeptides and acid-hydrolyzed dipeptides was examined. It was hypothesized that dipeptides should not trigger allosteric modulation of RNA structure, but that acid-mediated hydrolysis of dipeptides (Fig. 20C) carrying at least 1 lysyl residue should become active. As predicted, 179 *lysC* does not undergo allosteric modulation upon the addition of the dipeptides lys-lys, lys-ala, ala-lys, or ala-ala (Fig. 20D). However, the three dipeptides that carry at least one lysyl residue induce structural modulation of RNA upon pretreatment of the dipeptides with 6 N HCl at 115°C for 23 hr, followed by evaporation and neutralization. The extent of structural modulation (Fig. 20E) indicates that the samples containing the hydrolyzed lysine-containing dipeptides fully saturate the *lysC* aptamer, which is in accordance with the acid-mediated release of saturating amounts (greater than 1 μ M; see below) of L-lysine.

It was also observed that an intermediate level of structural modulation occurs when D-lysine is pre-treated with HCl. Interestingly, the published rate of epimerization between D- and L-lysine (Engel, M.H., and Hare, P.E. 1982. Racemization rates of the basic amino acids. *Year Book Carnegie Inst. Washington* 81:422-425) is sufficient to account for the approximately 1 μ M of L-lysine that is needed to produce half-maximal structural modulation (Fig. 20E). These results are consistent with lysine acting as the molecular ligand for the *lysC* aptamer, and that RNA conformational changes are not due to unknown contaminants of the commercial L-lysine preparation.

iii. Binding affinity and stoichiometry of the *B. subtilis* L-lysine aptamer

An approximation of the dissociation constant (K_D) was made by conducting in-line probing assays with 179 *lysC* using various concentrations of L-lysine (Fig. 21A). The sites of structural modulation exhibit progressively lower levels of spontaneous cleavage in response to increasing concentrations of ligand. A plot of the extent of RNA cleavage versus concentration of L-lysine (Fig. 21B) indicates that half-maximal structural modulation occurs when approximately 1 μ M amino acid is present in the mixture, thus reflecting the apparent K_D of the 179 *lysC* for its target ligand.

The apparent K_D value for a longer construct that encompasses structural elements predicted to be involved in transcription termination exhibits a significantly poorer affinity for L-lysine. Specifically, an RNA construct encompassing nucleotides 1

through 315 of the *lysC* 5'-UTR was found by in-line probing to exhibit an apparent K_D of $\sim 500 \mu\text{M}$. Similar differences in ligand affinities for other riboswitches have been observed, wherein the minimized aptamer binds more tightly its cognate ligand compared to the same aptamer in the context of the complete riboswitch (aptamer plus the adjoining expression platform). This is most likely due to the presence of competing secondary or tertiary structures that might be important for the function of the riboswitch as a genetic control element, but that reduce ligand binding affinity by reducing pre-organization of the aptamer domain.

Equilibrium dialysis also was used to examine the affinity and specificity of the 179 *lysC* aptamer for its target (Fig. 21C). In the absence of RNA, tritiated L-lysine is expected to distribute equally between the two chambers (*a* and *b*) of an equilibrium dialysis apparatus. However, the addition of excess aptamer to one chamber of the system should shift the distribution of tritium towards this chamber as a result of complex formation. This asymmetric distribution of tritium is expected to be restored to unity by the addition of a large excess of unlabeled competitor ligand, which displaces the bulk of the tritiated lysine from the RNA. As expected, the fraction of tritiated L-lysine in chamber *b* of the equilibrium dialysis apparatus is ~ 0.5 in the absence of RNA (Fig. 21C) after a 10 hr incubation. This fraction is altered to ~ 0.8 after incubation when a 200-fold excess of 179 *lysC* ($10 \mu\text{M}$) is added to chamber *b*, while this symmetric distribution of tritium is restored upon incubation for an additional 10 hours after the introduction of excess ($50 \mu\text{M}$) unlabeled L-lysine. Furthermore, D-lysine and L-ornithine do not restore equal distribution of tritium, which is consistent with their failure to modulate RNA structure as determined by in-line probing.

A Scatchard plot also was created by using data from a series of equilibrium dialysis experiments conducted under various concentrations of tritiated L-lysine (Fig. 21D). The slope of the resulting line indicates that the 179 *lysC* RNA binds to L-lysine with an apparent K_D of $\sim 1 \mu\text{M}$, which is consistent with that observed by using in-line probing. Furthermore, the x intercept of the line occurs near an *r* value of 1, which demonstrates that the RNA forms a 1:1 complex with its ligand.

iv. The lysine aptamer and adjacent sequences function as an amino acid-dependent riboswitch

With a number of riboswitches examined to date, there is a discernable set of structures residing immediately downstream of the aptamer domain that serve to control

gene expression in response to ligand binding. Typically, the structure of this “expression platform” is modulated by metabolite binding to the aptamer domain. The alternative structure subsequently leads to modulation of transcription or translation processes. For example, the TPP riboswitch on the *thiM* mRNA of *E. coli* carries an expression platform that appears to preclude ribosome binding to the Shine-Dalgarno sequence of the adjacent coding region (see Example 2). Similarly, the expression platforms of various riboswitches from *B. subtilis* undergo ligand-induced formation of a stem-loop structure that induces transcription termination (e.g. Examples 3, 6 and 7).

It has been reported that the *lysC* mRNA undergoes transcription termination in cultured *B. subtilis* cells grown in the presence of excess L-lysine (Kochhar, S., and Paulus, H. 1996, *Microbiol.* 142:1635-1639.). It was observed herein that a sequence domain that participates in forming the P1 stem of the *lysC* aptamer is complementary to a portion of the putative terminator hairpin that resides ~30 nucleotides downstream (Fig. 22A). This architecture is similar to that of several other riboswitches, some of which exhibit termination of transcription *in vitro* upon addition of the corresponding ligand as cited above. Therefore, the *lysC* leader sequence appears to serve as a L-lysine-specific riboswitch that induces transcription termination by modulating the formation of a terminator stem.

In vitro transcription assays were conducted in the absence and presence of L-lysine and several analogs (Fig. 22B, left). In the absence of added ligand, single-round transcription *in vitro* using *E. coli* RNA polymerase produces terminated product corresponding to ~36% of the total transcription yield. In contrast, the amount of terminated product increases to ~76% when 10 mM L-lysine is present during *in vitro* transcription. Neither D-lysine nor L-ornithine induce termination, which is consistent with the fact that these compounds are not recognized by the lysine aptamer domain and thus are not expected to trigger transcription termination.

The configuration of the expression platform for the *lysC* gene in *B. subtilis* strongly implicates a transcription termination mechanism, wherein the binding of L-lysine is expected to stabilize the P1 stem, thus permitting formation of the terminator hairpin (Fig. 22A). This proposed mechanism was examined by placing mutations within the critical pairing elements and by assessing lysine-induced transcription termination (Fig. 22B, center). Specifically, variant M1 carries two mutations that disrupt the formation of the terminator stem. This variant loses lysine-dependent modulation of

transcription termination, and produces greater transcriptional read-through relative to the wild-type construct. M2 carries a total of four mutations that compensate for the disruption of the terminator stem, but that cause disruption of the anti-terminator stem. This construct also loses lysine-dependent modulation, whereas the amount of the
5 terminated product expectedly becomes greater. Finally, the six-nucleotide variant M3 that carries the same mutations as M2 plus two additional mutations to restore the anti-terminator base-pairing potential results in near wild-type performance with regards to lysine-mediated modulation of transcription termination. These findings are consistent with a riboswitch mechanism wherein lysine binding precludes formation of an anti-terminator stem, thus increasing transcription termination by formation of an intrinsic
10 terminator structure.

v. Evidence that Riboswitches Serve as Antibiotics Targets

Unlike other lysine analogs, both L-lysine hydroxymate and the antimicrobial compound thiosine (*S*-(2-aminoethyl)-L-cysteine; Fig. 22A, inset) cause an increase in
15 transcription termination (Fig. 22B, left). These two compounds exhibit the best apparent K_D values of any of the analogs tested, with values for L-lysine hydroxymate and thiosine of ~100 μ M and ~30 μ M, respectively (data not shown). In previous studies, a series of mutants were identified in *B. subtilis* (Vold, B., et al., 1975, *J. Bacteriol.* 121:970-974; Lu, Y., et al., 1992, *FEMS Microbiol. Lett.* 92:23-27) and *E. coli* (Patte, J.-C., et al.,
20 1998, *FEMS Microbiol. Lett.* 169:165-170) that cause resistance to thiosine and cause derepression of *lysC* expression. These mutations all map to the lysine aptamer domain (see Fig. 22A for select *B. subtilis* mutants), and all appear to cause disruptions in the conserved elements or the base-pairing integrity of the structure.

The functional integrity of two thiosine-resistant mutants (G39A and G40A) was
25 examined by equilibrium dialysis and by in line probing, and both mutants fail to exhibit lysine-binding activity. Furthermore, RNA constructs that carry mutations in the otherwise conserved P1-P2 junction fail to undergo lysine-dependent transcription termination *in vitro* (Fig. 22B, right). These findings suggest that the antimicrobial action of thiosine might at least partially be due to direct binding of the analog to the lysine
30 riboswitch, causing repression of aspartokinase expression to a level that is deleterious to cell growth.

The function of the wild-type 5'-UTR of *lysC* and of the two thiosine-resistant mutants were also examined *in vivo* by fusion to a *lacZ* reporter gene. The wild-type riboswitch domain exhibits ligand-dependent modulation upon addition of L-lysine, whereas the G39A and G40A mutants fail to regulate β -galactosidase expression (Fig. 22C, medium II versus III). In contrast, lysine hydroxymate fails to repress expression of the reporter gene *in vivo* (medium IV), indicating that this compound might not attain a sufficiently high concentration inside cells to trigger transcription termination. As with lysine, thiosine also represses β -galactosidase expression for the wild-type construct, but not the two derepression mutants (medium V). This latter observation is consistent with the antimicrobial action of thiosine being due largely to its function as an effector for the lysine riboswitch.

3. Conclusions

The first mutants that caused deregulation of lysine biosynthesis in *B. subtilis* were identified nearly three decades ago (Vold, B., et al., 1975, *J. Bacteriol.* 121:970-974), however, the mechanism of genetic regulation has remained unresolved. Disclosed herein, it was demonstrated that the 5'-UTR of the *lysC* mRNA from *B. subtilis* serves as a riboswitch that responds to the amino acid lysine. The derepressed mutants isolated in the original study cause disruption of the aptamer domain of the riboswitch, such that the ligand is no longer bound by the RNA. Furthermore, *in vivo* expression studies using mutant *lysC* fragment-reporter gene fusions indicate that these riboswitch mutations most likely cause unregulated over-expression of aspartokinase, which catalyzes the first step in the biosynthetic pathway to lysine and several other amino acids.

Bacteria use various mechanisms to respond genetically to amino acid concentrations. Two of the more prominent mechanisms, translation-mediated transcription attenuation and T box-dependent mechanisms (Henkin, T.M., and Yanofsky, C. 2002, *BioEssays* 24:700-707), both sense the presence of non-aminoacylated tRNAs. Indeed, 18 of the 20 common amino acids in *B. subtilis* appear to be detected indirectly through the use of T box elements. Interestingly, there is no known tRNA^{lys}-dependent T-box in any organism, and presumably the lysine riboswitch described herein serves as the genetic sensor for this amino acid in the absence of a corresponding T box. Moreover, the genetic distribution of lysine riboswitches affiliated

with the *nhaC* gene from several organisms indicates that this RNA genetic element might be a key regulator of cellular pH.

Since the *lysC* mRNA functions as receptor for L-lysine, the Lys riboswitch can serve as a drug target. (See other examples, Hesselberth, J.R., and Ellington, A.D. 2002, *Nature Struct. Biol.* 9:891-893; Sudarsan, N., et al., 2003, *RNA* 9:644-647). The lysine riboswitch, and perhaps other classes of riboswitches as well, can be targeted by analogs that selectively bind to the riboswitch and induce genetic modulation. In *B. subtilis*, an analog of lysine that triggers the riboswitch would be expected to function as an antimicrobial agent, because the reduction of aspartokinase expression should induce starvation for lysine and other critical metabolites. The finding that thiosine binds to the lysine aptamer *in vitro*, and causes down regulation of a reporter construct fused to the wild-type riboswitch, provides support for the view that riboswitches are a newly recognized class of targets for drug discovery.

Recent discoveries have been elucidating the roles of small RNAs in guiding gene expression in a wide range of organisms (for a review see Gottesman, S. 2002, *Genes Dev.* 16:2829-2842). It is apparent that small RNAs, including riboswitch domains embedded within mRNAs, can control gene expression by a wide range of mechanisms. Unlike other RNA genetic control elements, riboswitches directly bind to metabolites and control the expression of genes that are involved in the import and biosynthesis of a number of fundamental metabolites. Riboswitches examined previously respond to compounds that are chemically related to nucleotides. However, the existence of a class of riboswitches that responds to a small amino acid with high selectivity serves as proof that natural RNA switches can detect and respond to a greater range of metabolite classes.

F. Example 6: Guanine and Other Riboswitches in *Bacillus subtilis* and Other Bacteria

1. Summary

Riboswitches are metabolite-binding domains within certain messenger RNAs that serve as precision sensors for their corresponding targets. Allosteric rearrangement of mRNA structure is mediated by ligand binding, and this results in modulation of gene expression. A class of riboswitches that selectively recognizes guanine and becomes saturated at concentrations as low as 5 nM are disclosed herein. In *Bacillus subtilis*, this mRNA motif is located on at least five separate transcriptional units that together encode

17 genes that are mostly involved in purine transport and purine nucleotide biosynthesis. These findings provide further examples of mRNAs that sense metabolites and that control gene expression without the need for protein factors. Furthermore, it is now apparent that riboswitches contribute to the regulation of numerous fundamental metabolic pathways in certain bacteria.

2. Introduction

It is widely understood that the interplay of protein factors and nucleic acids guide the complex regulatory networks for genetic expression in modern cells. In most instances, protein factors appear to be well-suited agents for maintaining genetic expression networks. Proteins can adopt complex shapes and carry out a variety of functions that permit living systems to sense accurately their chemical and physical environments. Protein factors that respond to metabolites typically act by binding DNA to modulate transcription initiation (e.g. the lac repressor protein; Matthews, K.S., and Nichols, J.C., 1998, *Prog. Nucleic Acids Res. Mol. Biol.* 58, 127-164) or by binding RNA to control either transcription termination (e.g. the PyrR protein; Switzer, R.L., et al., 1999, *Prog. Nucleic Acids Res. Mol. Biol.* 62, 329-367) or translation (e.g. the TRAP protein; Babitzke, P., and Gollnick, P., 2001, *J. Bacteriol.* 183, 5795-5802). Protein factors respond to environmental stimuli by various mechanisms such as allosteric modulation or post-translational modification, and are adept at exploiting these mechanisms to serve as highly responsive genetic switches (e.g. see Ptashne, M., and Gann, A. (2002). *Genes and Signals*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY).

In addition to the widespread participation of protein factors in genetic control, it is also known that RNA can take an active role in genetic regulation. Recent studies have begun to reveal the substantial role that small non-coding RNAs play in selectively targeting mRNAs for destruction, which results in down-regulation of gene expression (e.g. see Hannon, G.J. 2002, *Nature* 418, 244-251 and references therein). This process of RNA interference takes advantage of the ability of short RNAs to recognize the intended mRNA target selectively via Watson-Crick base complementation, after which the bound mRNAs are destroyed by the action of proteins. RNAs are ideal agents for molecular recognition in this system because it is far easier to generate new target-specific RNA factors through evolutionary processes than it would be to generate protein factors with novel but highly specific RNA binding sites.

Many studies have now confirmed that the complex three-dimensional shapes that some RNA molecules can mimic protein receptors and antibodies in their ability to selectively bind proteins or even small molecules (Gold, L., et al., 1995, *Annu. Rev. Biochem.* 64, 763-797; Hermann, T., and Patel, D., 2000, *Science* 287, 820-825).

5 Furthermore, RNAs exhibit sufficient structural complexity to permit the formation of allosteric domains that undergo structural and functional modulation upon ligand binding (Soukup, G.A., and Breaker, R.R., 1999a, *Proc. Natl. Acad. Sci. USA* 96, 3584-3589; Seetharaman, S. et al., 2001, *Nature Biotechnol.* 19, 336-341). Natural RNAs also are capable of binding nucleotides, as demonstrated by the group I self-splicing RNA, which
10 binds guanosine or its phosphorylated derivatives (McConnell, T.S., et al., 1993, *Proc. Natl. Acad. Sci. USA* 90, 8362-8366). More recently, evidence has been provided which indicates that direct binding of ATP by an RNA is essential for packaging DNA into a viral capsid (Shu, D., and Guo, P., 2003, *J. Biol. Chem.* 278, 7119-7125).

The known riboswitches bind their target metabolites with high affinity and
15 precision, which are essential characteristics for any type of molecular switch that can permit accurate and sensitive genetic control. For example, a recently identified riboswitch that responds to the coenzyme *S*-adenosylmethionine (SAM) binds it target with a dissociation constant (K_D) of ~4 nM (see Example 7). Furthermore, the riboswitch can discriminate ~100-fold against *S*-adenosylhomocysteine, which is a natural
20 metabolite that differs from SAM by a single methyl group and an associated positive charge. Disclosed herein (Example 1) genetic control involving riboswitches is a widespread phenomenon with regard to its biological distribution and the target molecules that are being monitored. The observations that certain mRNAs from Archaeal organisms carry riboswitch-like domains (Stormo, G.D., and Ji., Y., 2001, *Proc. Natl. Acad. Sci. USA* 98, 9465-9467; Rodionov, D.A., et al., 2002, *J. Biol. Chem.* 277, 48949-
25 48959) and that several mRNAs from fungi and plants bind thiamine pyrophosphate (TPP) (Sudarsan, N., et al., 2003, *RNA* 9:644-647).

The genetic regulation of purine transport and purine biosynthesis pathways in bacteria, which are fundamental to the metabolic maintenance of nucleotides and nucleic
30 acids (Switzer, R.L., et al., 2002, A.L. Sonenshein, *et al.*, eds., ASM Press, Washington, pp. 255-269), were analyzed for the presence of riboswitches. In *B. subtilis*, numerous genes are involved in the biosynthesis of purines (*pur* operon with 12 genes; Ebbole, D.J., and Zalkin, H. 1987, *J. Biol. Chem.* 262, 8274-8287) and in the salvage of purine

bases from degraded nucleic acids. The involvement of a regulatory protein factor has been proposed to participate in the control of the *xpt-pbuX* operon that encodes a xanthine phosphoribosyltransferase and a xanthine-specific purine permease, respectively (Christiansen, L.C., et al., 1997, J. Bacteriol. 179, 2540-2550). Although the protein factor PurR is known to serve as a repressor of transcription in the presence of elevated adenine concentrations (Weng, M., et al., 1995, Proc. Natl. Acad. Sci. USA 92, 7455-7459), no protein with corresponding function has been identified in *B. subtilis* that responds to guanine.

Disclosed herein the *xpt-pbuX* operon is controlled by a riboswitch that exhibits high affinity and high selectivity for guanine. This newfound class of riboswitches is present in the 5'-untranslated region (5'-UTR) of five transcriptional units in *B. subtilis*, including that of the 12-gene *pur* operon. Thus, direct binding of guanine by mRNAs serves as a critical determinant of metabolic homeostasis for purine metabolism in certain bacteria. Furthermore, it was determined that the known classes of riboswitches, which respond to seven distinct target molecules, appear to control at least 68 genes in *Bacillus subtilis* that are of fundamental importance to central metabolic pathways. These findings indicate that riboswitches play a substantial role in metabolic regulation in living systems that direct interaction between small metabolites and RNA is a significant and widespread form of genetic regulation in bacteria.

3. Experimental Procedures

i. Chemicals and Oligonucleotides

Guanine and its analogs xanthine, hypoxanthine, adenine, guanosine, 7-methylguanine, N²-methylguanine, 1-methylxanthine, 3-methylxanthine, 8-methylxanthine, 2-aminopurine, 2,6-diaminopurine, allopurinol, 2-amino-6-mercaptapurine, lumazine, and guanine-8-³H hydrochloride were purchased from Sigma. Inosine, uric acid, 2-amino-6-bromopurine, O-methyl guanine and pterin were purchased from Aldrich.

DNA oligonucleotides were synthesized by the Keck Foundation Biotechnology Resource Center at Yale University, purified by denaturing PAGE and eluted from the gel by crush-soaking in 10 mM Tris-HCl (pH 7.5 at 23°C), 200 mM NaCl, and 1 mM EDTA. Oligonucleotides were recovered from solution by precipitation with ethanol.

ii. Phylogenetic Analyses

G box domains were identified by sequence similarity to the *xpt-pbuX* 5'-UTR by conducting a BLASTN search of Genbank using default parameters. These hits were expanded by searching for degenerate matches to the pattern (<<<< [2] TA [6] <<< [2] 5 ATNNGG [2] >>> [5] GTNTCTAC [3] <<<<< [3] CCNNNAA [3] >>>>> [5] >>>>). Angled brackets indicate base pairing. Bracketed numbers are variable gaps with constrained maximum lengths denoted. A total of four violations of this pattern were permitted when forming the phylogeny depicted in Figure 23. It is important in this instance to note that only the BS3-*xpt* domain (that of the *xpt-pbuX* leader) has been 10 shown to bind guanine. It was demonstrated that the molecular specificity of the VV1 representative is for adenine and not guanine (unpublished data). Given the possible trivial means by which a guanine-binding RNA aptamer might be altered to bind adenine (e.g. a C to U change if the C residue is used by the aptamer to make a Watson-Crick-pairing interaction with guanine), it cannot be ruled out that other representatives also 15 have altered molecular recognition.

iii. In-line Probing of RNA Constructs

The *B. subtilis* 201 *xpt* leader and truncated 93 *xpt* aptamer RNAs were prepared by *in vitro* transcription using T7 RNA polymerase and the appropriate PCR DNA templates, and were subsequently 5' ³²P-labeled using a protocol similar to that described 20 previously (Seetharaman, S. et al., 2001, Nature Biotechnol. 19, 336-341). Labeled precursor RNAs (~2 nM) were subjected to in-line probing using conditions similar to those described in Example 2. Reactions (10 µL) were incubated for 40 hr at 25°C in a buffer containing 50 mM Tris (pH 8.5 at 25°C), 20 mM MgCl₂ and 100 mM KCl in the presence or absence of purines as indicated for each experiment. Purine concentrations 25 ranging from 1 nM to 10 µM were typically employed but ranged as high as 300 µM for poor-binding ligands. Denaturing 10% PAGE was used to separate spontaneous cleavage products and a Molecular Dynamics PhosphorImager was used to view the results. Quantitation of spontaneous cleavage yields was achieved by using ImageQuaNT software. Since concentrations of RNA below 2 nM for in-line probing cannot be used 30 easily due to insufficient levels of signal, apparent *K_D* values near this concentration reflect the maximum possible value.

iv. Equilibrium Dialysis

Equilibrium dialysis assays were conducted using a DispoEquilibrium Dialyzer (ED-1, Harvard Bioscience), wherein two chambers *a* and *b* were separated by a 5,000 MWCO membrane. The final composition of buffer included 50 mM Tris-HCl (pH 8.5 at 25°C), 20 mM MgCl₂ and 100 mM KCl (30 µL delivered to each chamber). Chamber *a* also contained 100 nM ³H-guanine, while chamber *b* also contained 300 nM of *xpt* RNA constructs as indicated for each experiment. After 10 hr of equilibration at 25°C, a 5 µl aliquot from each chamber was removed for quantitation by liquid scintillation counter. When appropriate, an additional 5 µL of buffer was added to *a* and an equivalent volume of buffer containing 500 nM unlabeled purine was added to *b*. After an additional 10 hr incubation at 25°C, 5 µl aliquots were again drawn for quantitation of tritium distribution.

v. Construction of *xpt-lacZ* Fusions

Genetic manipulations were conducted using approaches similar to those described elsewhere herein. Briefly, a DNA construct encompassing nt -121 to +197 relative to the transcription start site of the *xpt-pbuX* operon from *B. subtilis* strain 1A40 (Bacillus Genetic Stock Center, Columbus, OH) was PCR amplified as an *Eco*R1-*Bam*H1 fragment. The product was cloned into pDG1661 at a site directly upstream of the *lacZ* reporter gene. Mutants were created within the engineered pDG1661 by using the appropriate primers and the QuickChange Site-directed mutagenesis kit (Stratagene). Plasmid variants were integrated into the *amyE* locus of strain 1A40. Transformants were selected for chloramphenicol (5 µg ml⁻¹) resistance and screened for sensitivity to spectinomycin (100 µg ml⁻¹). The integrity of each construct was confirmed by sequencing.

vi. Guanine-mediated Modulation of β-galactosidase Expression

B. subtilis cells were grown with shaking at 37°C in minimal media containing 0.4% w/v glucose, 20 g L⁻¹ (NH₄)₂SO₄, 25 g L⁻¹ K₂HPO₄·3H₂O, 6 g L⁻¹ KH₂PO₄, 1 g L⁻¹ sodium citrate, 0.2 g L⁻¹ MgSO₄·7H₂O, 0.2% glutamate, 5 µg ml⁻¹ chloramphenicol, 50 µg ml⁻¹ L-tryptophan, 50 µg ml⁻¹ L-lysine and 50 µg ml⁻¹ L-methionine. Purines were added at a final concentration of 0.5 mg ml⁻¹. Cells at mid exponential stage (A₅₉₅ of ~0.1) were harvested by centrifugation and resuspended in minimal media in the absence or presence of a purine (0.5 mg mL⁻¹) as indicated for each experiment. Although the

poor solubility of guanine causes the formation of a detectable level of precipitate at this concentration, no adverse affects of cell growth were observed. Unless otherwise specified, cells were incubated for an additional 3 hrs before performing β -galactosidase assays. Data presented in Figure 28C was generated as described above with the exception that β -galactosidase assays were performed at the times indicated.

4. Results and Discussion

i. A Conserved Domain in the 5'-UTR of Several *B. subtilis* mRNAs.

The *xpt-pbuX* operon is regulated by guanine, hypoxanthine, and xanthine. These purine compounds share chemical similarity and are adjacent to each other in the pathways of purine salvage. In contrast to the *pur* operon, regulation of the *xpt-pbuX* operon remains unaffected by adenine in a strain wherein adenine deaminase is inactive (Christiansen, L.C., et al., 1997, J. Bacteriol. 179, 2540-2550). These observations had fostered speculation that an unidentified protein factor might be involved in guanine recognition (Ebbole, D.J., and Zalkin, H. 1987, J. Biol. Chem. 262, 8274-8287), however, such a genetic factor has not been identified. Moreover, the 5'-UTR of the *xpt-pbuX* mRNA is rather large (185 nucleotides), which could be sufficient to accommodate a riboswitch domain.

Riboswitches are typically composed of two functional domains: an aptamer that selectively binds its target metabolite and an expression platform that responds to metabolite binding and controls gene expression by allosteric means. The most conserved portion of known riboswitches is the aptamer domain, whereas the adjoining expression platform can vary widely in sequence and in secondary structure. The high sequence conservation of the aptamer is due to the fact that the RNA must retain its ability to form a receptor for a chemical that does not change through evolution. In contrast, the expression platform can form one of a great diversity of structures that permit genetic control in response to ligand binding by the aptamer domain. This evolutionary conservation was exploited to conduct a database search for *xpt-pbuX* 5'-UTR sequences that are present in other *B. subtilis* genes and also in other bacterial species. Five transcriptional units within *B. subtilis* that closely correspond in sequence and predicted secondary structure with nucleotides 14 through 82 of the *xpt-pbuX* 5'-UTR (Figure 23) were identified. A total of 32 representatives of this domain were

identified amongst several Gram-positive and Gram-negative bacteria. Other members can exist as well.

From this representative set of RNAs, a consensus sequence and secondary structure for the conserved RNA motif termed the "G box" (Figure 24A) were identified.

5 The secondary structure of the G box is composed of a three-stem (P1 through P3) junction, wherein significant sequence conservation occurs within P1 and in the unpaired regions. Furthermore, it was found that stems P2 and P3 both favor seven base pairs in length with one- or two-base mismatches permitted. This unusual conservation of stem length implies that these structural elements establish distance and orientation constraints of their stem-loop sequences relative to the three-stem junction. Some base-pairing potential exists between the two stem-loop sequences, which might permit the formation of a pseudoknot. These characteristics indicate that G-box domains most likely use conserved secondary- and tertiary-structure elements to adopt a precise three-dimensional fold.

15 ii. **The G box RNA from the *xpt-pbuX* 5'-UTR of *B. subtilis* Binds Guanine.**

Two RNA constructs based on the *xpt-pbuX* 5'-UTR of *B. subtilis* were prepared to examine whether the mRNA selectively binds guanine or its closest analogs. A double-stranded DNA template corresponding to the entire 5' UTR and the first four codons of the *xpt-pbuX* mRNA was generated by PCR using primers that introduced a promoter sequence for T7 RNA polymerase and several nucleotide additions and mutations that permit further manipulation (Figure 24B; see also Experimental Procedures). A truncated form of this construct also was created by PCR that encompasses the 5' half of the UTR. Upon transcription, the shorter DNA template generates a 93-nucleotide transcript termed 93 *xpt*, while the longer template produces a 201-nucleotide transcript termed 201 *xpt*.

These precursor RNAs were 5' ³²P-labeled and subjected to an in-line probing assay (e.g. see Example 1) wherein the spontaneous cleavage of RNA linkages within an aptamer is monitored in the presence and absence of its corresponding ligand. It was found that the patterns of spontaneous cleavage of the 93 *xpt* (Figure 24C) and the 201 *xpt* (Figure 25A) RNAs undergo significant alteration upon addition of guanine at a concentration of 1 μ M. Both hypoxanthine and xanthine also induce modulation of spontaneous cleavage at this concentration. Specifically, four major regions exhibit ligand-mediated reduction in spontaneous cleavage (Figure 24B and 24C). However, the

presence of 1 μ M adenine (and as much as 1 mM) does not alter the pattern of RNA cleavage products. These results indicate that the G box domain in the 5' UTR of the *B. subtilis xpt-pbuX* mRNA serves as an aptamer for guanine and related purines, and that this aptamer undergoes significant structural modulation upon ligand binding. In the context of a riboswitch, this allosteric function could be harnessed by the mRNA to modulate structural elements that regulate gene expression.

In a preliminary assessment of the affinity that the guanine aptamer has for its target, in-line probing with 201 *xpt* in the presence of various concentrations of guanine was conducted. As expected, increasing concentrations provided progressively decreasing amounts of spontaneous cleavage at the four major sites of structural modulation (Figure 25A). Half-maximum levels of modulation were observed when a concentration of ~ 5 nM guanine is used for in-line probing (Figure 25B). Although this implies that the K_D for 201 *xpt* under these conditions is ~ 5 nM, it is important to note that the actual value might be somewhat lower because of the limitations of the in-line probing assay (see Experimental Procedures). In addition, the K_D was determined under non-physiological conditions (e.g. high Mg^{2+} and elevated pH), and so the binding affinity might be somewhat different *in vivo*. However, using this number for comparison, the affinity of the 201 *xpt* RNA for guanine is more than 10,000-fold greater than that of the *Tetrahymena* group I ribozyme for its guanosine monophosphate substrate (McConnell, T.S., et al., 1993, Proc. Natl. Acad. Sci. USA 90, 8362-8366). This difference most likely reflects the relative differences in concentrations of the two compounds that the RNAs experience inside their respective cellular environments.

iii. The Guanine Aptamer Discriminates Against Many Purine Analogs

To maintain precise metabolic homeostasis, the cell must be able to sense the concentration of its target metabolite, but also must prevent regulatory cross talk with other compounds that otherwise might inadvertently trigger genetic modulation. Indeed, a hallmark of other riboswitches is the ability to discriminate between closely related metabolites. For example, the FMN and TPP riboswitches discriminate against the unphosphorylated coenzyme precursors thiamine and riboflavin by $\sim 1,000$ fold (see Examples 2 and 3).

This requirement for obligate molecular discrimination against related metabolites is expected to be extreme with guanine riboswitches, as there are numerous purine nucleosides and nucleotides, purine bases, and purine-like compounds that are

present in the cell. Using the in-line probing strategy described in Figure 25, the apparent K_D values of the 93 *xpt* RNA were established for a variety of purines and purine analogs. Hypoxanthine and xanthine exhibit K_D values that are closest to the value determined for guanine, while adenine has a K_D value in excess of 300 μ M (Figure 26A).

5 These results are consistent with the observation that adenine does not significantly repress expression of the *xpt-pbuX* operon as do the other purines (Christiansen, L.C., et al., 1997, *J. Bacteriol.* 179, 2540-2550). However, it is not clear whether hypoxanthine and xanthine might repress gene expression by directly binding a guanine riboswitch, or whether they might first be converted into guanine before influencing genetic control.

10 It was found that alteration of every functionalized position on the guanine heterocycle causes a substantial loss of binding affinity (Figure 26B, Figure 27). For example, the oxygen atom at position 6 of guanine is a significant determinant of molecular recognition, as demonstrated by the losses in apparent K_D for 2-aminopurine (>10,000-fold loss), 2-amino-6-bromopurine (~1,000 fold), and O⁶-methylguanine (>100
15 fold). Most molecular interactions could be explained by invoking hydrogen-bonding contacts between the RNA and guanine with the exception of the molecular interaction at C8. Here, presumably the RNA structure creates a steric clash with analogs that carry additional bulk, such as 8-methylxanthine (>10,000 fold) and uric acid (>10,000 fold).

A summary of the likely molecular recognition features that the guanine aptamer
20 requires for maximum affinity is depicted in Figure 26C. However, the likely possibility that significant binding affinity could be derived through base stacking was not examined. The presence of so many productive contacts between the RNA and all faces of guanine suggest that the ligand is most likely entirely engulfed by the aptamer's structure. This would also explain why the RNA is capable of generating recognition via
25 steric occlusion of bulkier compounds such as uric acid. In certain biological environments, for example, uric acid can build up to high concentrations that permit crystallization. In such environments, a bacterium would require a high level of discrimination to prevent undesirable repression of guanine-regulated genes. In light of such molecular recognition challenges, it is not surprising that an RNA genetic switch
30 would evolve extensive molecular contacts with its target compound.

iv. Confirmation of Guanine Aptamer Function by Equilibrium Dialysis

Equilibrium dialysis was used to provide further evidence that the G box RNA from the *xpt-pbuX* operon binds guanine preferentially over other purines and purine

analogs. A substantial shift in tritiated guanine is expected to occur in a two-chamber dialysis apparatus when an excess of functional RNA is added to one chamber (Figure 27A). Furthermore, this shifted equilibrium should return to unity upon addition of an excess of unlabeled competitor ligand. As expected, it was observed that greater than
5 90% of tritiated guanine co-localizes with 93 *xpt* RNA, and subsequently redistributes when an excess of unlabeled guanine is introduced. In contrast, the presence of excess unlabeled analogs has no effect on co-localization of ³H-guanine and the RNA (Figure 27B). Even the nucleoside guanosine (9-ribosylguanine) fails to restore equal distribution of guanine between the two chambers, which is consistent with the RNA folding to form
10 a tight pocket for the base alone.

Both in-line probing and equilibrium dialysis data indicate that this natural aptamer binds guanine with high affinity and specificity. In a previous study, *in vitro* evolution was used to isolate a purine-binding aptamer from a pool of random-sequence RNAs (Kiga, D., et al., 1998, Nucleic Acids Res. 26, 1755-1760). This engineered
15 aptamer exhibits a K_D of 1.3 μ M for guanine, and shows only a 2- to 3-fold discrimination against hypoxanthine and xanthine. The lower specificity and affinity of this aptamer for selected purines is due to the fact that only the N1, N7 and O6 positions are important for molecular recognition. In contrast, the G box RNA appears to make productive contacts with all available functional groups on guanine, presumably through
20 hydrogen bonding (Figure 26C).

v. Aptamer Mutations Affect Guanine Binding and Genetic Control

A variety of mutations were introduced into the G box domain to examine the importance of several structural elements and conserved nucleotides (Figure 28A). The influence of these mutations on guanine binding was determined in the context of the 93
25 *xpt* RNA by using equilibrium dialysis. Mutations that independently disrupt the three stems (M1, M4 and M6) cause a loss of binding function, as does a variant RNA (M3) that carries two mutations in the central junction (Figure 28B). In contrast, the effects of the disruptive stem mutations are largely reversed by making compensatory mutations (M2, M5 and M7) that restore base pairing. These results are consistent with the
30 phylogenetic analysis (Figure 23), which indicates that stem structure is important but that the precise sequence composition of these elements is of less importance.

Binding function of variant aptamers *in vitro* also correlates with genetic control *in vivo*. The results disclosed herein confirmed earlier findings that a reporter gene

carrying the 5'-UTR of the *xpt-pbuX* mRNA is repressed by guanine, and to a lesser extent by hypoxanthine and xanthine (Christiansen, L.C., et al., 1997, J. Bacteriol. 179, 2540-2550). Specifically, transcriptional fusions were created between a β -galactosidase reporter gene and variant *xpt-pbuX* 5'-UTR sequences carrying the mutations described in Figure 28A. *B. subtilis* chromosomal transformants using the wild-type sequence exhibit the expected levels of genetic modulation (Figure 28C). Although the *xtp* aptamer exhibits dissociation constants for xanthine and hypoxanthine that are essentially identical *in vitro*, the differences in genetic modulation by these compounds *in vivo* might be due to differences in their cellular concentrations.

Aptamer variants with impaired guanine binding *in vitro* also exhibit a loss of β -galactosidase repression (Figure 28D). Furthermore, restoration of base pairing in stems P1 through P3 results in restored genetic control. The M2 variant is of particular interest because it not only exhibits restored genetic control, but also provides modest expression of β -galactosidase in the absence of guanine. Riboswitch function requires the action of an aptamer for molecular sensing as well as an expression platform that transduces RNA-ligand complex formation into a genetic response. Examples of TPP and FMN riboswitches (see Examples 2 and 3) appear to function by differential formation of terminator and antiterminator structures. Such ligand-induced formation of transcription anti-termination structures also appears to be the basis of expression platform mechanisms used by numerous SAM riboswitches (see Example 7). Construct M2 carries three mutations within the putative anti-terminator structure of the *xpt-pbuX* leader, and thus is expected to exhibit an overall reduction of reporter gene expression because these mutations should bias structure folding towards terminator stem formation.

The results of these mutational and functional analyses confirm the major features of the secondary structure model (P1 through P3) and demonstrate that they are critical for metabolite binding. Furthermore, the correlation between ligand binding and genetic control indicates that the G box and adjacent nucleotides of the *xpt-pbuX* leader sequence operate in concert to function as a guanine-dependent riboswitch, most likely by operating via allosteric control of transcription termination.

vi. Riboswitches Control Fundamental Biochemical Pathways

Our findings indicate that the G box RNA of the *xpt-pbuX* operon is a key structural element of a guanine-sensing riboswitch that exhibits extraordinary affinity

and selectivity for its target. In *B. subtilis*, this general riboswitch motif appears to control at least five transcriptional units (Figure 23). Although the precise function of several of the gene products in this newly identified regulon have not been clearly defined, the known genes from *B. subtilis* and from other organisms are mostly related to purine metabolism. Based on the results disclosed herein, it is likely the G box domain within the 5'-UTR of this large *pur* operon is responsible for guanine-dependent riboswitch regulation, and that the genetic regulatory mechanism might be similar to that proposed herein for the *xpt-pbuX* operon.

The distribution of G box domains in *B. subtilis* and other bacteria suggests that this class of metabolite-binding RNAs controls a regulon that is essential for cell survival. In *B. subtilis*, guanine riboswitches (or related adenine-dependent riboswitches – see the legend to Figure 23) appear to provide at least some contribution to the genetic regulation of 17 genes. The discovery of guanine-dependent riboswitches adds to a growing list of similar metabolite-sensing RNAs. For example, a class of riboswitches that responds to SAM (McDaniel, B.A.M., et al., 2003, Proc. Natl. Acad. Sci. USA 100, 3083-3088; Epshtein, V., et al., 2003, Proc. Natl. Acad. Sci. USA 100, 5052-5056) controls a regulon of as many as 26 genes that are involved in coenzyme biosynthesis, amino acid metabolism, and sulfur metabolism. When included with genes that are controlled by other riboswitch classes, at least 68 genes (nearly 2% of its total genetic complement) are under riboswitch control (Figure 29).

Riboswitches for ligands such as guanine and SAM apparently are serving as master control molecules whose concentrations are being monitored to ensure homeostasis of a much wider set of metabolic pathways. Riboswitches also seem to permit metabolite surveillance and genetic control with the same level of precision and efficiency as that exhibited by protein factors. Therefore, these RNA switches could have emerged late in the evolution of modern biochemical architectures because they are functionally comparable to genetic switches made of protein. However, given their fundamental role in metabolic maintenance and the widespread phylogenetic distribution of certain riboswitches, it is consistent that aptamer domains similar to these might have been the primary mechanism by which RNA-world organisms detected metabolites and controlled biochemical pathways before the emergence of proteins.

5. Conclusions

This demonstration that guanine is sensed by metabolite-binding mRNAs expands the known classes of riboswitches, and provides additional evidence that certain bacterial RNAs are responsible for monitoring the concentrations of critical coenzymes and other compounds that are fundamental to all living systems. Phylogenetic analyses and biochemical data indicate that many bacteria and, in some instances, eukaryotes (Sudarsan, N., et al., 2003, *RNA* 9:644-647) entrust riboswitches to sense essential metabolites and mediate genetic control. Although protein factors undoubtedly could be used to carry out these important regulatory tasks, based on the disclosure herein, highly structured RNAs are well suited for this role. If RNA polymers were a poorly suited medium for generating metabolite receptors with high affinity and precision, then one would expect that evolution would have long ago replaced them by protein factors.

Disclosed herein it is consistent (e.g. see Examples 1 and 2) that riboswitches are derivatives of an ancient genetic control system that monitored metabolic and environmental signals before the evolutionary emergence of proteins. Interestingly, each of the metabolite targets of riboswitches has been proposed to come from an RNA world (White, H.B. III., 1976, *J. Mol. Evol.* 7, 101-104; Benner, S.A., et al., 1989, *Proc. Natl. Acad. Sci. USA* 86, 7054-7058; Jeffares, D.C., et al., 1998, *J. Mol. Evol.* 46, 18-36; Jadhav, V.R., and Yarus, M., 2002, *Biochimie* 84, 877-888). The identification of guanine as a trigger for riboswitches is consistent with metabolite sensing RNAs having originated very early in evolution. Also disclosed herein is another class of riboswitches that responds to the amino acid lysine (Figure 29). Although all riboswitches could be more recent evolutionary inventions, even the origin of the lysine riboswitch might date from before the last common ancestor and back to a time when living systems were transitioning from a pure RNA world to a more modern metabolic state that made use of encoded protein synthesis.

G. Example 7: *S*-adenosylmethionine Riboswitches

Riboswitches are metabolite-binding RNA structures that serve as genetic control elements for certain messenger RNAs. These RNA switches have been identified in all three kingdoms of life and are typically responsible for the control of genes whose protein products are involved in the biosynthesis, transport, or utilization of the target metabolite. Disclosed herein, is a highly conserved RNA domain found in bacteria serves as a riboswitch that responds to the coenzyme *S*-adenosylmethionine (SAM) with

remarkably high affinity and specificity. SAM riboswitches undergo structural reorganization upon introduction of SAM, and these allosteric changes regulate the expression of 26 genes in *Bacillus subtilis*. This and related findings indicate that direct interaction between small metabolites and allosteric mRNAs is a significant and widespread form of genetic regulation in bacteria.

1. Results

i. Identification of a SAM-responsive riboswitch

Each of the compounds sensed by previously identified riboswitches (coenzyme B₁₂, TPP, FMN) is used as a coenzyme by modern protein enzymes. Interestingly, these coenzymes have significant structural similarity to RNA, which has been used to support speculation that they might also have been used as coenzymes by ancient ribozymes in an RNA world (S. A. Benner, et al., *Proc. Natl. Acad. Sci. USA* **86**, 7054 (1989); H. B. White III, *J. Mol. Evol.* **7**, 101 (1976); D. C. Jeffares, et al., *J. Mol. Evol.* **46**, 18 (1998). If modern riboswitches are direct descendants of RNA control systems that originated in the RNA world, then the metabolites they sense and the metabolic pathways that they control will be of fundamental importance to modern biochemical processes. To further assess this hypothesis, a search for additional riboswitches, to determine their biochemical characteristics, and to establish their role in genetic control on a genome-wide level was performed.

In this effort the S box was examined (F. J. Grundy, T. M. Henkin, *Mol. Microbiol.* **30**, 737 (1998)), which is a highly conserved sequence domain (Fig. 30A) that is located within the 5'-untranslated region (5'-UTR) of certain messenger RNAs in Gram-positive bacteria. Both genetic and sequence analyses suggest that the S box domain serves as a genetic control element for a regulon composed of 11 transcriptional units. These mRNAs encode as many as 26 different genes in *B. subtilis* that are involved in sulfur metabolism, methionine biosynthesis, cysteine biosynthesis, and SAM biosynthesis. However, the nature of the putative regulatory factor and the metabolite to which it responds had not been established (T. M. Henkin, *Curr. Opin. Microbiol.* **3**, 149 (2000); F. J. Grundy, T. M. Henkin, *Frontiers Biosci.* **8**, D20 (2003)). An RNA construct corresponding to the first 251 nucleotides of the *yitJ* mRNA of *B. subtilis* (Fig. 30b) was prepared by *in vitro* transcription (G. A. Soukup, R. R. Breaker, *RNA* **5**, 1308 (1999)). The *yitJ* gene product is a putative methylene tetrahydrofolate reductase – an enzyme proposed to be involved in methionine biosynthesis (F. J. Grundy, T. M. Henkin,

Mol. Microbiol. **30**, 737 (1998). The 251 *yitJ* RNA was subjected to “in-line probing”, which reveals locations of structured and unstructured portions of RNA polymers by relying on the variability in rates of spontaneous RNA phosphodiester cleavage caused by differences in structural context. In-line probing can also reveal nucleotides
5 participating in metabolite-induced structural modulation (see Examples 1-3).

Whether the 251 *yitJ* RNA might bind *S*-adenosylmethionine (SAM) was analyzed. Indeed, upon separation by polyacrylamide gel electrophoresis (PAGE), the pattern of spontaneous RNA cleavage products (Fig. 30c) was indicative of a highly structured RNA element that undergoes conformational modulation upon introduction of
10 SAM to a final concentration of either 0.1 mM or 1 mM. In contrast, no structural modulation was evident upon the introduction of methionine at the same concentrations, suggesting that the RNA might require both the methionine and 5'-deoxyadenosyl moieties of SAM to induce structural reorganization. The locations of the ligand-induced modulations (Fig. 30b) indicated that the conserved core of the S box RNA serves as a
15 natural aptamer (L. Gold, et al., *Annu. Rev. Biochem.* **64**, 763 (1995)). for SAM. Similar results were observed with 124 *yitJ*, which encompasses nucleotides 28 through 149 of the mRNA leader plus two G residues at the 5' terminus.

ii. Molecular recognition by a SAM-dependent riboswitch

A genetic switch that responds to metabolites must be able to bind its target with
20 a dissociation constant (K_D) that is relevant to physiological concentrations. Furthermore, the metabolite receptor must be able to discriminate precisely against closely related compounds that are likely to occur in the same milieu, or risk undesirable modulation of gene expression. Therefore, the affinity of the *yitJ* RNA for SAM was assessed, and the ability of the RNA to discriminate against biologically relevant compounds that are
25 structurally similar to this target (Fig. 31a).

The K_D of 251 *yitJ* for SAM was determined by using in-line probing to monitor the extent of structural modulation over a range of ligand concentrations (Fig. 31b, left). Although the K_D of 251 *yitJ* for SAM is ~200 nM, the minimized aptamer domain represented by 124 *yitJ* exhibits a K_D of ~4 nM under the disclosed assay conditions.
30 Such improvements in binding affinity by minimized aptamer domains have been observed (see Example 2). This most likely reflects greater structural preorganization of the ligand binding form of the aptamer domain due to the elimination of the adjoining expression platform, which otherwise would permit alternative folding to occur. Tight

binding was also observed when the 124 *yitJ* was interrogated by using a Scatchard analysis with tritiated SAM. The assessment of binding affinity indicated that the K_D for the 124 *yitJ* aptamer is more than 1000-fold improved compared to that reported recently for a related RNA (McDaniel, B. et al., *Proc. Natl. Acad. Sci. USA* **100**, 3083-3088 (2003)). Normal concentrations of SAM in bacteria are typically in the low micromolar range (McDaniel, B. et al., *Proc. Natl. Acad. Sci. USA* **100**, 3083-3088 (2003)), however, most of this coenzyme pool is probably bound by enzymes. Therefore the low K_D exhibited by this riboswitch might be needed to sense the concentration of free SAM.

As expected, the 124 *yitJ* RNA achieves a high level of molecular discrimination against analogs of SAM. For example, the RNA exhibits ~100-fold discrimination against SAH (Fig. 31*b*, right), which is produced upon utilization of SAM as a coenzyme for methylation reactions (F. Takusagawa, et al., In: *Comprehensive Biological Catalysis*, M. Sinnott, ed., Academic Press, Vol. 1, pp. 1-30 (1998)). Thus, the aptamer must form a binding pocket for SAM that can sense the absence of a single methyl group and an associated loss of positive charge. Similarly, the RNA discriminates nearly 10,000 fold against SAC, which is another biological compound that differs from SAH by the absence of a single methylene group. This pattern of molecular discrimination was confirmed by using equilibrium dialysis (Fig. 31*c*).

iii. SAM binding by an mRNA is required for genetic regulation

The secondary structure model for the SAM-binding aptamer domain was established using phylogenetic data (F. J. Grundy, T. M. Henkin, *Mol. Microbiol.* **30**, 737 (1998)). To provide further support for this model, the influence of disruptive and compensatory mutations (Fig. 32*a*) on the binding function of the 124 *yitJ* RNA, and on SAM-mediated genetic control of a *lacZ* reporter gene when fused with variant riboswitches based on these mutant aptamers was examined. Mutations that alter the conserved core of the aptamer (M1) or that disrupt base pairing in each of the four major base-paired regions (M2, M4, M6 and M8) largely result in a loss of SAM binding function as determined by equilibrium dialysis (Fig. 32*b*). Compensatory mutations that restore base pairing in these stems (M3, M5, M7, M9) restore at least partial binding activity.

It has been shown (F. J. Grundy, T. M. Henkin, *Mol. Microbiol.* **30**, 737 (1998)) that a growth medium rich in methionine leads to repression of *B. subtilis* genes that carry the S box domain. This is most likely due to the ability of the cell to convert

methionine into an ample supply of SAM. Disclosed herein in all cases tested, the binding function of the mutant correlates with their ability to down regulate an appended reporter gene when presented with excess methionine in otherwise minimal growth media (Fig. 32c). These findings are consistent with SAM binding to the mRNA being
5 necessary for the genetic regulation of S box mRNAs.

**iv. SAM riboswitches control gene expression by transcription termination
in *B. subtilis***

Disclosed herein bacterial riboswitches can control gene expression by modulating either transcription termination or translation initiation (see Examples 2 and
10 3), while several putative riboswitches in eukaryotes might use one of several different mechanisms. In *B. subtilis*, the SAM-binding aptamer domains typically reside immediately upstream from a putative transcription terminator hairpin (F. J. Grundy, T. M. Henkin, *Mol. Microbiol.* **30**, 737 (1998)), which implies that SAM binding most likely induces transcription termination as described previously for FMN- and TPP-
15 dependent riboswitches (see Example 3).

In vitro transcription in the absence or presence of SAM using 11 DNA templates corresponding to the mRNA leader sequences of the S box regulon was performed. These assays were simplified by using T7 RNA polymerase instead of the native *B. subtilis* RNA polymerase. It was observed that an FMN-dependent riboswitch induces
20 transcription termination even when T7 RNA polymerase is used as a surrogate for the bacterial polymerase (see Example 3). In this study, it was found that the *yitJ*, *yoaD* and *metK* leader constructs exhibit modest transcription termination upon the addition of SAM. More dramatically, the termination product from the *metI* leader construct increases from ~12% to nearly 75% upon introduction of SAM (Fig. 33a). In all
25 instances, little or no modulation of transcription termination occurs when the analogs SAH or SAC are added to the reaction. The remaining seven S-box representatives did not exhibit significant modulation with T7 RNA polymerase, presumably because it serves as an imperfect substitute for the native polymerase. Indeed, SAM-dependent transcription termination is observed with many of these mRNA leader sequences when
30 *E. coli* or *B. subtilis* polymerases are used in the assay (McDaniel, B. et al., *Proc. Natl. Acad. Sci. USA* **100**, 3083-3088 (2003)).

The mechanism of SAM-induced termination (Fig. 33b) most likely involves the ligand-mediated formation of alternative hairpin structures that permit transcriptional

read-through (anti-terminator formation without SAM) or that cause termination (terminator formation with SAM). This mechanism was examined by generating several mutant *metI* constructs that carry disruptive or compensatory changes in the expression platform (Fig. 33b). SAM causes an additional ~20% yield in transcription termination in a mutant (*Mabc*) that carries six mutations relative to the wild-type *metI* riboswitch, which retains proper terminator and anti-terminator base complementation. However, incomplete representation of these six mutations that do not permit normal pairing interactions to occur permits little or no SAM-mediated transcription modulation. Furthermore, mutations that disrupt terminator stem formation (*Ma*) yield lower levels of termination, while mutations that disrupt anti-terminator stem formation (*Mab*, *Mc*) yield higher levels of termination (Fig. 33b). These findings indicate that the RNA structural modulation induced by SAM binding mediates genetic control by sequestering an anti-terminator sequence, and thus favors the formation of a transcriptional terminator hairpin.

v. Riboswitches control multiple genes that are involved in fundamental biochemical pathways

The 11 transcriptional units that comprise the regulon controlled by SAM riboswitches (F. J. Grundy, T. M. Henkin, *Mol. Microbiol.* **30**, 737 (1998)) appear to encompass at least 26 genes that are central to sulfur metabolism, amino acid metabolism, and SAM biosynthesis. Although all 11 transcriptional units from *B. subtilis* carry a consensus S box element, a recent report indicates that gene expression from one of these (*cysH*) is not modulated by addition of methionine to the medium, as are other S box RNAs (M. C. Mansilla, et al., *J. Bacteriol.* **182**, 5885 (2000)). The aptamer domain from *B. subtilis* *cysH* does bind SAM with an affinity that is more than 2 orders of magnitude poorer than that of *yitJ* from the same organism (Fig. 34a). However, the *cysH* homolog from *B. anthracis* exhibits a K_D that matches that of *yitJ* (Fig. 34b), implying that the *B. subtilis* *cysH* aptamer has suffered one or more mutations that have somewhat degraded binding affinity.

2. Conclusion

Current biochemical and bioinformatics data indicate that *B. subtilis* has at least 68 genes (nearly 2% of its total genetic complement) under riboswitch control. Moreover, each of these mRNAs is responding to biological compounds that are universal in biology. The fact that genetic control elements for fundamental metabolic

processes are formed by RNA indicates that this polymer has the structural sophistication needed to precisely monitor chemical environments and transduce metabolite binding events into genetic responses. A more detailed analysis of riboswitch structures at the atomic level would be of great utility in determining how metabolite binding promotes allosteric reorganization RNA genetic switches.

Riboswitches for ligands such as SAM and guanine appear to be serving as master control molecules whose concentrations are being monitored to ensure homeostasis of a much wider set of metabolic pathways. Riboswitches seem to permit metabolite surveillance and genetic control with the same level of precision and efficiency as that exhibited by protein factors, and thus could have emerged late in the evolution of modern biochemical architectures.

3. Methods

i. DNA oligonucleotides and chemicals

Synthetic DNAs were purchased from The Keck Foundation Biotechnology Resource Center at Yale University. Preparation of RNAs by *in vitro* transcription was conducted (Seetharaman, S., et al., *Nat. Biotechnol.* 19, 336-341 (2001)) and the products were purified as described in Example 2. SAM, various analogs of SAM, and *S*-adenosyl-L-methionine-methyl-³H (³H-SAM) were purchased from Sigma.

ii. DNA constructs

A *yitJ* DNA construct encompassing nucleotides -380 to +15 relative to the translation start site was prepared using primers that generated *EcoRI* and *BamHI* restriction sites upon PCR amplification of *B. subtilis* chromosomal DNA (strain 168). The product was cloned into pDG1661 (ref. 26; *Bacillus* Genetic Stock Center, Columbus, OH) using these restriction sites, which places the riboswitch immediately upstream of the *lacZ* reporter gene. Mutants were created by using the appropriate mutagenic primers and the QuickChange site-directed mutagenesis kit (Stratagene). All sequences were confirmed by sequencing.

iii. *In vivo* analysis of riboswitch function

B. subtilis strain 1A234 was obtained from the *Bacillus* Genetic Stock Center, Columbus, OH. Cells were grown with shaking at 37°C either in rich media (2XYT broth or tryptose blood agar base) or defined media (0.5% w/v glucose, 20 g L⁻¹ (NH₄)₂SO₄, 183 g L⁻¹ K₂HPO₄·3H₂O, 60 g L⁻¹ KH₂PO₄, 10 g L⁻¹ sodium citrate, 2 g L⁻¹ MgSO₄·7H₂O, 5 μM MgCl₂, 50 μg L⁻¹ tryptophan, and 50 μg L⁻¹ glutamate. Methionine

was added to $50 \mu\text{g L}^{-1}$ for routine growth. Growth under methionine-limiting conditions was established by incubation under routine growth conditions to an A_{595} of 0.1, at which time the cells were pelleted by centrifugation, resuspended in minimal media, split into two aliquots, and supplemented with either $50 \mu\text{g L}^{-1}$ (+ methionine) or $0.75 \mu\text{g L}^{-1}$ (- methionine) (Fig. 32c). Cultures were incubated for an additional 3 hr before performing β -galactosidase assays. Transformations of pDG1661 variants (see DNA constructs) into *B. subtilis* were performed as described elsewhere (H. Jarmer, et al., *FEMS Microbiol. Lett.* **206**, 197 (2002)). The correct transformants were identified by selecting for chloramphenicol ($5 \mu\text{g mL}^{-1}$) resistance and screening for spectinomycin ($100 \mu\text{g mL}^{-1}$) sensitivity. Proper site-specific genomic insertion by double cross-over recombination was confirmed by PCR using *amyE*-specific primers.

iv. *In vitro* transcription termination assays

Transcription reactions (10 μL) containing ~30 pmoles of specific template DNA, 200 μM each NTP, 5 μCi [α - ^{32}P]UTP (1 Ci = 37 GBq) and 50 units of T7 RNA polymerase (New England Biolabs) were incubated in the presence of 50 mM Tris-HCl (pH 7.5 at 23°C), 15 mM MgCl_2 , 2 mM spermidine, 5 mM DTT at 37°C for 2 hr. SAM and its analogs were added to a final concentration of 50 μM . Transcription templates were generated for all 11 riboswitch domains in the S box regulon of *B. subtilis* by using PCR with corresponding primers that in each case produced transcripts beginning with GG, encompassing the putative natural transcription start (F. J. Grundy, T. M. Henkin, *Mol. Microbiol.* **30**, 737 (1998)), and including the first 13 codons of the adjoining open reading frame. Transcription products were separated by denaturing 6% PAGE and visualized by PhosphorImager. Termination yields were approximated by determining the ratio of RNAs in the termination band relative to the combined terminated and full-length RNAs.

H. Example 8: Adenine Riboswitches

A class of riboswitches that recognizes guanine and discriminates against most other purine analogs was recently identified (see Example 6). Representative RNAs that carry the consensus sequence and structural features of guanine riboswitches are located in the 5'-untranslated region (UTR) of numerous genes of prokaryotes, where they control expression of proteins involved in purine salvage and biosynthesis. This example shows that three representatives of this phylogenetic collection bind adenine with values for apparent dissociation constant (apparent K_D) that are several orders of magnitude

better than for guanine. The preference for adenine is due to a single nucleotide substitution in the core of the riboswitch, wherein each representative most likely recognizes its corresponding ligand by forming a Watson/Crick base pair. In addition, the adenine-specific riboswitch associated with the *ydhL* gene of *Bacillus subtilis* functions as a genetic 'ON' switch, wherein adenine binding causes a structural rearrangement that precludes formation of an intrinsic transcription terminator stem.

Guanine-sensing riboswitches are a class of RNA genetic control elements that modulate gene expression in response to changing concentrations of this compound (see Example 6). This is one of a number of classes of metabolite-binding riboswitches that regulate gene expression in response to various fundamental compounds such as lysine and the coenzymes FMN, SAM, B₁₂ and TPP (thiamin pyrophosphate) (see Example 6). Typically, each riboswitch is composed of two functional domains, an aptamer and an expression platform, that function together as a transducer of chemical signals into altered patterns of gene expression. The aptamer serves as a specific receptor for the target metabolite, wherein ligand binding brings about allosteric changes in both the aptamer and expression platform domains.

Detailed examinations of the ligand specificities for the natural aptamers from guanine- and lysine-specific riboswitches have been conducted (see Example 6), and less comprehensive examinations of the FMN, SAM, B₁₂ and TPP aptamers have been conducted (see Examples 1-3). In each case, the RNAs exhibit high levels of molecular discrimination by disfavoring the binding of even closely related metabolite analogs. This characteristic of high molecular discrimination is a hallmark of enzymes and receptors, including genetic regulatory factors, which need to carry out biological processes with great precision in the presence of complex chemical mixtures.

The molecular recognition characteristics of guanine riboswitches are distinguished by the fact that nearly every position around the purine heterocycle appears to be critical for high affinity binding by the aptamer. Thus, the arrangement of the binding pocket permits the riboswitch to control gene expression in response to changing guanine concentrations, but prevents modulation of gene expression in response to increasing concentrations of adenine (see Example 6; Christiansen, L.C., et al., *J. Bacteriol.* 179, 2540-1550 (1997)). However, it is likely that receptors made of RNA, like their protein counterparts, could acquire altered molecular recognition characteristics as a result of natural selection. This would permit riboswitches to emerge through

evolution that selectively sense and respond to metabolites that are proximal in metabolic pathways.

This example confirms the existence of a variant class of riboswitches that responds to adenine. These riboswitches carry an aptamer domain that corresponds closely in sequence and secondary structure to the guanine aptamer described recently (see Example 6). However, each representative of the adenine sub-class of riboswitches carries a C to U mutation in the conserved core of the aptamer, indicating that this residue is involved in metabolite recognition. The results indicate that the identity of this single nucleotide determines the binding specificity between guanine and adenine, which provides an example of how complex riboswitch structures could mutate to recognize new metabolite targets.

1. RESULTS

i. Phylogenetic comparison between riboswitch domains

A comparative sequence strategy was used to identify a series of intergenic regions from a number of prokaryotic species that carry a conserved sequence element termed the "G box" (see Example 6). *B. subtilis* carries at least five of these motifs, which were also identified using genetics techniques (Johansen, L.E., et al., *J. Bacteriol.* 185, 5200-5209). Each representative of the phylogeny has three potential base-paired elements (P1 through P3) and as many as 24 nucleotides that are conserved in greater than 90% of the examples identified to date. A subset of this phylogeny with features common to the G box motif highlighted is presented herein (Fig. 35A). When selected representatives are examined in greater detail, they are encompassed by the mRNA transcript of the gene immediately downstream, and thus are present as RNA elements located in the 5'-UTR of certain mRNAs.

Several notable differences present in the guanine-binding domain of *xpt* (Fig. 35B) relative to the RNA from *ydhL* (Fig. 35C) were identified. First, among the 23 sequence variations in *ydhL* compared to *xpt*, 20 reside within base-paired elements and most of these changes permit base pairing to be retained. This strongly indicates that the overall secondary structure between the two RNAs is similar. Second, the remaining three mutations reside in unpaired regions, such that two (corresponding to positions 31 and 48 relative to *xpt*) reside at locations that are known to be variable. These mutations do not impact significantly the structure and function of the RNA. Third, the remaining mutation is a C to U change at position 74 relative to *xpt*, which otherwise corresponds to

a strictly conserved nucleotide of the three-stem junction. Given the location of this mutation, this change might alter the molecular recognition characteristics of the *ydhL* aptamer.

ii. Variant G box RNAs selectively bind adenine

5 It had been established (see Example 6) that the *xpt* aptamer makes numerous contacts with its ligand, and that as many as seven hydrogen bonds might be involved in forming the RNA-ligand complex. Furthermore, there is evidence that steric clashes also likely aid in restricting the range of metabolites that can be bound by the RNA. This array of contacts can only be established by forming multiple interactions between the
10 various sides of guanine and distal parts of the RNA.

An intriguing hypothesis is the possibility that the C residue at position 74 of *xpt* could conceivably be forming a Watson/Crick base pair with guanine, thus forming three of these hydrogen bonds. Since a U mutation resides in the corresponding position in *B. subtilis ydhL* and two RNAs from *C. perfringens* and *V. vulnificus*, we believe that these
15 RNAs might serve as adenine-responsive riboswitches. This hypothesis was further supported by recognition that the latter two genes (*add*) encode adenine deaminase enzymes. It seems reasonable that adenine should be the metabolite whose concentration is being monitored to determine the expression levels of adenine deaminase.

The ligand specificity of five G box RNAs (Fig. 35A) was examined by using in-line probing (Soukup, G.A. & Breaker, R.R., *RNA* 5, 1308-1325 (1999); Soukup, G.A., DeRose, E.C., *RNA* 7, 524-536 (2001)). In this assay, the spontaneous cleavage of RNA is monitored in the absence of ligand, or in the presence of guanine or adenine. As
20 predicted previously (see Example 6), the *purE* RNA (Fig. 36A) exhibits changes in the pattern of spontaneous cleavage products in the presence of guanine that correspond to that observed for the *xpt* RNA (Fig. 36B). These results confirm that the *purE* RNA, like
25 the *xpt* RNA, responds allosterically to guanine and not to adenine when incubated in the presence of the concentrations of ligand tested.

In contrast, all three RNAs that carry the C to U mutation in the junction between P1 and P3 (corresponding to C74 of *xpt*) do not respond to guanine, but exhibit structural
30 modulation only when incubated in the presence of adenine. Furthermore, the patterns of spontaneous cleavage for the adenine-specific aptamers are consistent with the secondary-structure model proposed for G box RNAs (Fig. 35). These results indicate that certain variants of the G box class of RNAs serve as sensors of adenine.

Furthermore, these findings are consistent with the hypothesis that, when located in their natural settings, the *ydhL* RNA from *B. subtilis* and the two *add* RNAs from *C. perfringens* and *V. vulnificus* serve as adenine-specific riboswitches.

iii. The *ydhL* aptamer binds adenine with high affinity and selectivity

5 Another characteristic of riboswitches is the aptamer domains exhibit tight binding for their corresponding target compound, and they discriminate against analogs, in some cases, by orders of magnitude in apparent K_D . For example, the guanine riboswitch from *B. subtilis xpt* exhibits an apparent K_D for guanine of ~5 nM, but binds adenine with an apparent K_D that is at least 100,000-fold poorer. In-line probing assays
10 were used to determine the binding affinities of the *B. subtilis* 80 *ydhL* RNA for these two purines. As expected, the RNA exhibits progressively changing patterns of spontaneous RNA cleavage fragments in the presence of increasing concentrations of adenine (Fig. 37A), but the pattern remains unchanged with increasing guanine concentrations as high as 10 μ M (see below).

15 The bands corresponding to spontaneous cleavage fragments that undergo change with increasing adenine concentrations were grouped into four sites and the extent of cleavage relative to the total RNA present were quantitated. This data was used to generate a plot (Fig. 37B) that provides an estimate of the apparent K_D for ligand binding. In this instance, half-maximal decrease in spontaneous cleavage at sites 1, 2 and
20 4, and the corresponding half-maximal increase in spontaneous cleavage at site 3 occurs when approximately 300 nM adenine is present in the in-line probing assay. Thus, the *ydhL* aptamer binds adenine with an apparent K_D that is similar to those exhibited by other classes of riboswitches.

The molecular recognition characteristics of 80 *ydhL* were further examined by
25 using the same in-line probing strategy with a variety of analogs. For example, a series of purine analogs that are close chemical variants to adenine exhibit measurable binding to the RNA (Fig. 38A). The ligands with measurable binding, 2,6-DAP, A and 2-AP, P, MA (listed in order of decreasing affinity), are all close analogs of adenine. Furthermore, the relative affinities of the RNA for various ligands provide some indication of the
30 contact points that the aptamer likely uses to establish molecular recognition (Fig. 38A, bottom right). This model is consistent with the finding that a series of purine analogs fail to exhibit measurable binding to the 80 *ydhL* RNA (Fig. 38B).

The collection of purines that are recognized by 80 *ydhL* indicate that only the Watson/Crick base-pairing face of the purine ligand is recognized differently by the *ydhL* aptamer compared to the *xpt* aptamer. For example, modification at the C8 position (8-chloroadenine) prevents ligand binding, which implies that a steric clash between certain purines and 80 *ydhL* as was observed for the *xpt* aptamer (see Example 6). Interestingly, the fact that 2,6-DAP, and not adenine, is the tightest-binding ligand provides insight into the similarities between the *ydhL* and *xpt* aptamers. This observation suggests that the 80 *ydhL* RNA retains at least one of the two hydrogen bond acceptor contacts that were proposed to exist in the *xpt* aptamer. Thus, the molecular recognition characteristics of these RNAs are consistent with the *ydhL* RNA differing in molecular recognition from *xpt* with a pattern that can be explained by a change from a Watson/Crick guanine-C base pair in *xpt* to a Watson/Crick adenine-U base pair in *ydhL*.

iv. Swapping ligand specificity of G box RNAs by molecular engineering

The idea that the *xpt* and *ydhL* RNAs might be deriving their specificity for guanine or adenine by a Watson/Crick base pairing interaction was examined in greater detail by using a molecular engineering approach. A similar approach was used previously (Wilson, K.S. & von Hippel, P.H. *Proc. Natl. Acad. Sci. USA* 92, 8793-8797) to change the ligand-rescue specificity of an abasic hammerhead ribozyme construct from guanine to adenine. Both wild-type (93 *xpt* and 80 *ydhL*) and mutant (93 *xpt* C to U and 80 *ydhL* U to C) forms of G box aptamers were generated and tested for binding activity with guanine and adenine (Fig. 39). The mutations correspond to nucleotide position 74 relative to the *xpt* sequence (Fig. 35B), which is suspected to be the determinant of molecular discrimination between guanine and adenine.

As observed previously (see Example 6), the aptamer based on *xpt* exhibits structural modulation only when incubated in the presence of guanine, and is able to shift the distribution of tritiated guanine (but not adenine) in an equilibrium dialysis assay (Fig. 39A). However, the 93 *xpt* RNA that carries a single C to U mutation at position 74 no longer is responsive to guanine, but exhibits structural modulation and binding activity during equilibrium dialysis only in the presence of adenine (Fig. 39B). In contrast, the wild-type 80 *ydhL* RNA is specific for adenine (Fig. 39C), while the corresponding U to C mutation at this critical nucleotide position alters binding specificity to guanine (Fig. 39D). Therefore, the primary determinant of the base specificity of G box aptamers is the C or U residue that is present in the junction between

stems P1 and P3, and that this base most likely forms a conventional Watson-Crick base pair with its target ligand.

v. Mechanism of genetic control by the *ydhL* adenine riboswitch from *B. subtilis*

5 In most instances, riboswitches control gene expression in prokaryotes by allosteric interconversion between alternate base-paired structures. For example, a TPP riboswitch from the *thiM* gene of *E. coli* makes use of alternate base pairing to sequester the Shine-Dalgarno sequence of the mRNA in the presence of ligand, presumably resulting in reduced translation initiation (see Example 2). In contrast, TPP riboswitches
10 from *B. subtilis* harness ligand-binding events to alter base-pairing patterns and form intrinsic terminator stems that cause transcription elongation to abort (Gusarov, I & Nudler, E. *Mol. Cell.* 4, 495-504 (1999); Mironov, A.S. *et al. Cell* 111, 747-756 (2002)). Similarly, metabolite-mediated formation of transcription terminator stems is a mechanism used by certain examples of riboswitches that respond to FMN (see Example
15 3 and 6), SAM (see Example 7), guanine (see Example 6), and lysine (see Example 5).

The UTR sequence of the *ydhL* riboswitch was examined to assess whether there is evidence of a transcription termination mechanism. Consistent with this possibility is the fact that the 5'-UTR of the *ydhL* mRNA can form a large hairpin, composed of as many as 22 base pairs, followed by a run of eight uridyl residues (Fig. 40A). This
20 structural feature, which was also noted elsewhere recently (Johansen, L.E., et al., *J. Bacteriol.* 185, 5200-5209), is characteristic of an intrinsic terminator stem. In the absence of adenine, it was considered that the riboswitch can form this intrinsic terminator. If true, then the genetic control status for this riboswitch would default to this predicted 'OFF' state, which prevents gene expression by inducing transcription
25 termination. In the presence of adenine, gene expression is expected to proceed because a substantial portion of the left shoulder of the terminator stem would be required to form stems P1 and P3 of the adenine aptamer domain. Since stems P1 and P2 are integral components of the adenine aptamer, ligand binding would establish a structure that precludes formation of the terminator stem.

30 This mechanism for the *ydhL* riboswitch was assessed *in vivo* by generating reporter constructs wherein various forms of guanine- and adenine-specific riboswitches were integrated into the *B. subtilis* genome. As controls, two reporter constructs were prepared with either the wild-type *xpt* riboswitch, or the *xpt* variant with the C to U

mutation at position 74. As expected, the wild-type *xpt* construct causes repression of β -galactosidase expression when presented with excess guanine in the culture medium (Fig. 40b). This finding is similar to those reported previously for function of the guanine riboswitch from *xpt* (see Example 6). Adenine also shows a modest (~4 fold) repression of reporter expression after a six-hour incubation. This latter effect is most likely due to the function of the PurR protein, which is known to provide modest down-regulation of transcription initiation in response to adenine at the *xpt-pbuX* promoter used in this construct (Cristiansen, L.C., et al., *J. Bacteriol.* 179, 2540-1550 (1997)).

A near identical *xpt* construct carrying the C to U mutation causes a loss of regulation upon addition of guanine, but shows no change in the putative protein-dependent control due to adenine (Fig. 40C). These results are consistent with the observed loss of guanine binding *in vitro* when this mutation is made, but suggest that the resulting specificity change to adenine *in vitro* does not permit robust adenine-dependent genetic control *in vivo*. Most likely, the diminished expression upon addition of adenine again is due to the PurR protein.

In contrast to the *xpt* riboswitch, the performance of the corresponding wild-type and mutant *ydhL* reporter constructs indicates that the latter is an adenine-dependent riboswitch with the opposite response to rising levels of ligand. Specifically, the wild-type *ydhL* construct exhibits very low β -galactosidase activity when assayed in the absence of ligand, or in the presence of guanine (Fig. 40D). However, a greater than 10-fold increase in gene expression occurs in response to added adenine. In addition, the single U to C mutation in the P1-P3 junction of the aptamer causes substantial (~100 fold) derepression regardless of what ligand is used (Fig. 40e). Although this seems counter to the model proposed for *ydhL* riboswitch function, it is important to note that this mutation indeed disrupts adenine binding, but it also causes a mismatch to occur in the terminator stem. If this mismatch is sufficiently destabilizing to the terminator stem, or if this mutation adversely affects the folding pathway for the riboswitch, then the default 'OFF' status for the genetic control element would be expected to change to default 'ON'. Therefore, the observed level of gene expression might be indicative of full activation of the *ydhL* gene when it's genetic control element is indifferent to the concentrations of purines in the cell.

2. DISCUSSION

i. The structure and evolution of adenine riboswitches

The sequence and biochemical similarities between guanine- and adenine-specific G box RNAs indicate that they are analogous in overall secondary and tertiary structure.

5 The ease of interchanging ligand specificities of these aptamers by making single mutations to the *xpt* and *ydhL* aptamers suggests that such changes might occur with high frequency in natural populations. However, the fact that neither single-base variant of the *xpt* or *ydhL* riboswitches exhibits corresponding specificity changes in genetic control *in vivo* suggests that multiple mutations might be necessary to make a useful swap in
10 riboswitch specificity.

It is important to note that the binding affinity of the resulting single-base *xpt* variant is not as robust for its new ligand. Specifically, the wild-type *xpt* RNA has an apparent K_D for guanine of no poorer than 5 nM (Fig. 39a), while the C to U variant of this RNA exhibits an apparent K_D for adenine of ~100 nM (Fig. 39b). In this case,
15 although the mutation results in a substantial change in base discrimination between guanine and adenine, binding affinity for the matched ligand has been somewhat degraded. In contrast, the wild-type and mutant *ydhL* RNAs exhibit both specificity change and retention of binding affinity for the matched ligands (Figs. 39C and 39D). However, the affinity for the U to C variant of 80 *ydhL* for guanine appears to be at least
20 10-fold poorer than that of 93 *xpt*.

Thus, accessory mutations that do not directly define ligand specificity but that further adjust the binding affinity might be necessary for G box RNAs to interconvert between guanine and adenine ligands in a biological setting. In this regard, it is interesting that the *ydhL* and *xpt* aptamers differ from each other at 23 positions (Fig.
25 35), with only one residing within an obviously critical position (C74 of *xpt*). Although some of these mutations might serve to fine-tune the binding affinity of the aptamers, many could be the result of neutral drift in the RNA sequence that is permitted because they retain the essential secondary-structure elements.

ii. Genetic control and function of the *ydhL* mRNA

30 Mutant strains of *B. subtilis* that resist the toxic effects of 2-fluoroadenine were reported recently (Johansen, L.E., et al., *J. Bacteriol.* 185, 5200-5209)). These mutations, which cause over-expression of the *ydhL* gene product, were mapped to the adenine riboswitch domain. In both instances, the changes (deletions) are expected to

disrupt riboswitch function by eliminating a portion of the terminator stem or by eliminating both the terminator stem and portions of the adenine aptamer domain. In both instances, the variants preclude the riboswitch from adopting its default state (transcription termination), which causes unmodulated activation of gene expression.

5 The protein product of the *ydhL* gene (also termed *pbuE*) has been proposed to be a purine efflux pump (Johansen, L.E., et al., *J. Bacteriol.* 185, 5200-5209)). Thus the resistance to 2-fluoroadenine conferred upon the cell by disruption of the adenine riboswitch from *ydhL* might be due to excretion of this toxic compound. In the natural genetic background, the presence of excess adenine within the cell most likely induces increased expression of the *ydhL* gene to produce the purine efflux protein. Higher levels
10 of this protein then work to normalize the concentration of purines by pumping out of the cell one or more forms of this compound class.

iii. Riboswitch mechanisms – genetic activation and deactivation by rising metabolite concentrations

15 The adenine riboswitch from *B. subtilis* also is notable for its mechanism of action. In the majority of riboswitches examined to date, metabolite binding causes a lowering of gene expression. This occurs either by ligand-mediated formation of a terminator stem to prevent transcription of the complete mRNA, or by sequestering the Shine-Dalgarno sequence and precluding translation initiation. In most instances, the
20 down-regulation of gene expression is expected, as a build-up of sufficient levels of a particular metabolite should logically provide a signal to turn off genes in that are involved in biosynthesis or import of the compound (Grundy, F.J. & Henkin, T.M. et al., *Frontiers Biosci.* 8, D20-31 (2003)).

 The adenine riboswitch from *ydhL* (and presumably for the *add* riboswitches as
25 well) belong to a group of genes whose functions would hint at the need for riboswitch activation in the presence of high concentrations of target compounds. In the case of *ydhL*, disposal of excess purines would seem to be an important capability given that certain purines such as guanine are insoluble at modest concentrations. Alternatively, there be no obvious need to express adenine deaminase if adenine concentrations were
30 exceptionally low, and therefore we expect that the riboswitches from the *add* genes of *C. perfringens* and *V. vulnificus* might be activated by ligand binding as well. Interestingly, T box domains, which are 5'-UTR structures that control the expression of many aminoacyl-tRNA synthetases in *B. subtilis* and other Gram-positive organisms

(Grundy, F.J., et al., *Proc. Natl. Acad. Sci. USA* 99, 11121-11126), also induce gene expression in response to rising concentrations of the target they sense. However, unlike the known metabolite-binding riboswitches, T box domains sense the biochemical precursor (non-aminoacylated tRNAs) to the products of the enzymes whose expression they control (Miller, J.H. *A Short Course in Bacterial Genetics*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York (1992)).

Although we expect that riboswitches that induce gene activation in response to increasing metabolite will occur less frequently due to genetic necessity, there is no inherent structural flaws in RNA folding that would skew this distribution between gene-activating and gene-deactivating riboswitches. Whether the riboswitch responds to ligand binding by activating or repressing gene expression, the RNAs will exploit allosteric changes in secondary and/or tertiary structure that are based on the same principles of RNA folding. The only obligate difference between activating and repressing riboswitches is in the fine structure of the expression platform, whereas the aptamer domain can remain largely unchanged.

3. METHODS

i. Purine analogs

Guanine, adenine, 2,6-diaminopurine, 2-aminopurine, hypoxanthine, xanthine, 1-methyladenine, purine, 6-methylaminopurine, N⁶-N⁶ dimethyladenine, 6-mercaptapurine, 3-methyladenine, guanine-8-³H and adenine-2,8-³H were purchased from Sigma. 6-cyanopurine and 8-azaadenine were obtained from Aldrich and 2-chloroadenine, 8-chloroadenine from Biolog Life Science Institute, Germany.

ii. DNA oligonucleotides

Oligonucleotides were synthesized by the HHMI Keck Foundation Biotechnology Resource Center at Yale University, purified by denaturing polyacrylamide gel electrophoresis, and were eluted from the gel by crush-soaking in a buffer containing 10 mM Tris-HCl (pH 7.5 at 23°C), 200 mM NaCl, and 1 mM EDTA. DNAs were precipitation with ethanol, resuspended in deionized water, and stored at -20°C until use.

iii. In-line probing of RNA constructs

RNA constructs were synthesized from the corresponding PCR DNA templates by transcription *in vitro* using T7 RNA polymerase, dephosphorylated, and 5'-end

labeled with ^{32}P as described in Example 6. In a typical in-line probing assay, 2 nM of labeled RNA were incubated in a buffer containing 20 mM MgCl_2 , 50 mM Tris-HCl (pH 8.3 at 25°C) and 100 mM KCl in the absence or presence of purine compounds as indicated for each experiment for 40 hrs at 25°C. Purine concentrations ranging from 1 nM to 10 μM were employed unless otherwise noted. At the end of each incubation, spontaneously cleaved products were separated on a denaturing (8 M urea) 10% PAGE, visualized using a PhosphorImager and quantitated using ImageQuaNT software (Molecular Dynamics).

iv. Equilibrium Dialysis

Equilibrium dialysis assays were conducted using a DispoEquilibrium Dialyzer (Harvard Biosciences), wherein chamber *A* and *B* are separated by a 5,000 MWCO membrane. Chamber *A* contained 30 μl of ^3H -guanine or ^3H -adenine at a concentration of 100 nM in a buffer containing 50 mM Tris-HCl (pH 8.5 at 25°C), 20 mM MgCl_2 , and 100 mM KCl. A 30 μl aliquot of the above mentioned buffer containing RNA at 3 μM concentration was delivered into chamber *B*. Equilibrations were allowed to proceed for 10 hrs at 25°C. Subsequently 5 μl was withdrawn from each chamber and quantitated by liquid scintillation counting.

v. Construction of *xpt*- and *ydhL-lacZ* fusions

A DNA construct encompassing nucleotides -468 to +9 relative to translational start site of *ydhL* was PCR amplified from *B. subtilis* strain 1A40 (Bacillus Genetic Stock Center, Columbus, OH) with primers that introduced EcoR1-BamH1 restriction sites. The wild-type construct was cloned into pDG1661 at EcoR1-BamH1 restriction sites directly upstream of the *lacZ* reporter gene and sequenced to confirm its integrity. The resulting plasmid was used as a template for site-directed mutagenesis via the QuickChange site-directed mutagenesis kit (Stratagene) using the appropriate primer. Plasmid variants were integrated into the *amyE* locus of *B. subtilis* strain 1A40 and the transformants were confirmed as described in Example 6.

vi. *In vivo* analysis of riboswitch function

Transformed *B. subtilis* cells were grown to mid log phase with constant shaking at 37°C in minimal media containing 0.4% w/v glucose, 20 g/l $(\text{NH}_4)_2\text{SO}_4$, 25 g/l K_2HPO_4 , 6 g/l KH_2PO_4 , 1 g/l sodium citrate, 0.2 g/L $\text{MgSO}_4 \cdot 7\text{H}_2\text{O}$, 0.2% glutamate, 5 $\mu\text{g/ml}$ chloramphenicol, 50 $\mu\text{g/ml}$ L-tryptophan, 50 $\mu\text{g/ml}$ L-lysine and 50 $\mu\text{g/ml}$ L-

methionine. Guanine or adenine was added to a final concentration of 0.1 mg/ml. Cells at mid exponential stage were harvested and resuspended in minimal media in the presence or absence of purines and grown for an additional time as indicated for each experiment, at which time 1 ml of cell culture was subjected to β -galactosidase activity assays using a variation of the method described by Miller (Miller, J.H. A Short Course in Bacterial Genetics. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York (1992)).

I. Example 9: Tables of sequence comparisons for the SAM, Cobalimin, Guanine, Adenine, and Lysine riboswitches discussed herein.

Figure 41 shows sequence and types of riboswitches. The alignment of these sequences is as disclosed herein, regions disclosed in the other figures correspond to the same regions in Figure 41.

Additional riboswitches were found based on published alignments and secondary structures (Grundy, F.J. & Henkin, T.M. The S box regulon: a new global transcription termination control system for methionine and cysteine biosynthesis genes in Gram-positive bacteria. *Mol. Microbiol.* **30**, 737-749 (1998)) using the SequenceSniffer program. This program finds degenerate matches to RNA patterns defined by linked sequence motifs and base pairing constraints. In the alignments, base pairing regions have the identical colored backgrounds and are labeled as in the corresponding figures discussed in Examples 1-8, with the addition of a putative pseudoknot marked PS. Predicted terminators (yellow) and start codons (green) are marked for some sequences. Positions for each sequence in the indicated Genbank record or unfinished genome contig are for the sequence column marked with a circle (•) – the fifth base in stem P1 that is 5' of the aptamer. Start is the offset from the column marked with an asterisk (*) – the sixth base in stem P1 that is 3' of the aptamer – to the start codon of the first gene in the operon. Genes were identified from COGNITOR (Tatusov, R.L., *et al.* The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* **29**, 22-28 (2001)) and PFAM (Bateman, A., *et al.* The Pfam Protein Families Database. *Nucleic Acids Res.* **30**, 276-280 (2002)) database matches to protein sequences annotated in the Genbank records. The standard names from these databases are used when possible (2011 = COG2011; ??? = no matches). Previous operon designations for *B. subtilis* are given in parentheses (Grundy, F.J. & Henkin, T.M. The S box regulon: a new global transcription

termination control system for methionine and cysteine biosynthesis genes in Gram-positive bacteria. *Mol. Microbiol.* 30, 737-749 (1998)). A subset of sequences with <90% pairwise identity between the bases encompassed by stem P1 was selected for determining the consensus sequence. In the consensus sequence, lowercase and
5 uppercase bases indicate >80% and >95% conservation at a position, respectively. Purine (R) and pyrimidine (Y) bases were assigned when no single base had >80% conservation.

(*) Sequence shares >90% identity with another sequence, and was excluded when determining the consensus.

10 (1) Very short hypothetical gene that may be a misannotated ORF.

(2) Possible S Box "pseudogene". The S Box is on the opposite strand 5' of the indicated operon.

It is understood that the disclosed method and compositions are not limited to the
15 particular methodology, protocols, and reagents described as these may vary. It is also to be understood that the terminology used herein is for the purpose of describing particular embodiments only, and is not intended to limit the scope of the present invention which will be limited only by the appended claims.

It must be noted that as used herein and in the appended claims, the singular
20 forms "a", "an", and "the" include plural reference unless the context clearly dictates otherwise. Thus, for example, reference to "a riboswitch" includes a plurality of such riboswitches, reference to "the riboswitch" is a reference to one or more riboswitches and equivalents thereof known to those skilled in the art, and so forth.

"Optional" or "optionally" means that the subsequently described event,
25 circumstance, or material may or may not occur or be present, and that the description includes instances where the event, circumstance, or material occurs or is present and instances where it does not occur or is not present.

Ranges may be expressed herein as from "about" one particular value, and/or to
"about" another particular value. When such a range is expressed, also specifically
30 contemplated and considered disclosed is the range from the one particular value and/or to the other particular value unless the context specifically indicates otherwise. Similarly, when values are expressed as approximations, by use of the antecedent "about," it will be understood that the particular value forms another, specifically

contemplated embodiment that should be considered disclosed unless the context specifically indicates otherwise. It will be further understood that the endpoints of each of the ranges are significant both in relation to the other endpoint, and independently of the other endpoint unless the context specifically indicates otherwise. Finally, it should
5 be understood that all of the individual values and sub-ranges of values contained within an explicitly disclosed range are also specifically contemplated and should be considered disclosed unless the context specifically indicates otherwise. The foregoing applies regardless of whether in particular cases some or all of these embodiments are explicitly disclosed.

10 Unless defined otherwise, all technical and scientific terms used herein have the same meanings as commonly understood by one of skill in the art to which the disclosed method and compositions belong. Although any methods and materials similar or equivalent to those described herein can be used in the practice or testing of the present method and compositions, the particularly useful methods, devices, and materials are as
15 described. Publications cited herein and the material for which they are cited are hereby specifically incorporated by reference. Nothing herein is to be construed as an admission that the present invention is not entitled to antedate such disclosure by virtue of prior invention. No admission is made that any reference constitutes prior art. The discussion of references states what their authors assert, and applicants reserve the right to challenge
20 the accuracy and pertinency of the cited documents. It will be clearly understood that, although a number of publications are referred to herein, such reference does not constitute an admission that any of these documents forms part of the common general knowledge in the art.

Throughout the description and claims of this specification, the word "comprise"
25 and variations of the word, such as "comprising" and "comprises," means "including but not limited to," and is not intended to exclude, for example, other additives, components, integers or steps.

Those skilled in the art will recognize, or be able to ascertain using no more than routine experimentation, many equivalents to the specific embodiments of the method
30 and compositions described herein. Such equivalents are intended to be encompassed by the following claims.

CLAIMS

We claim:

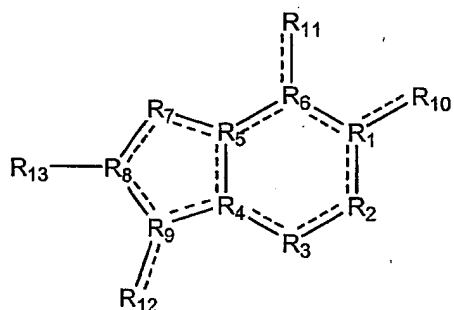
1. A regulatable gene expression construct comprising
a nucleic acid molecule encoding an RNA comprising a riboswitch operably linked to a coding region, wherein the riboswitch regulates expression of the RNA, wherein the riboswitch and coding region are heterologous.
2. The construct of claim 1 wherein the riboswitch comprises an aptamer domain and an expression platform domain, wherein the aptamer domain and the expression platform domain are heterologous.
3. The construct of claim 1 wherein the riboswitch comprises an aptamer domain and an expression platform domain, wherein the aptamer domain comprises a P1 stem, wherein the P1 stem comprises an aptamer strand and a control strand, wherein the expression platform domain comprises a regulated strand, wherein the regulated strand, the control strand, or both have been designed to form a stem structure.
4. A riboswitch, wherein the riboswitch is a non-natural derivative of a naturally-occurring riboswitch.
5. The riboswitch of claim 4 wherein the riboswitch comprises an aptamer domain and an expression platform domain, wherein the aptamer domain and the expression platform domain are heterologous.
6. The riboswitch of claim 4 wherein the riboswitch is derived from a naturally-occurring guanine-responsive riboswitch, adenine-responsive riboswitch, lysine-responsive riboswitch, thiamine pyrophosphate-responsive riboswitch, adenosylcobalamin-responsive riboswitch, flavin mononucleotide-responsive riboswitch, or a S-adenosylmethionine-responsive riboswitch.
7. The riboswitch of claim 4 wherein the riboswitch is activated by a trigger molecule, wherein the riboswitch produces a signal when activated by the trigger molecule.
8. A method of detecting a compound of interest, the method comprising
bringing into contact a sample and a riboswitch, wherein the riboswitch is activated by the compound of interest, wherein the riboswitch produces a signal when activated by the compound of interest, wherein the riboswitch produces a signal when the sample contains the compound of interest.

9. The method of claim 8 wherein the riboswitch changes conformation when activated by the compound of interest, wherein the change in conformation produces a signal via a conformation dependent label.

10. The method of claim 8 wherein the riboswitch changes conformation when activated by the compound of interest, wherein the change in conformation causes a change in expression of an RNA linked to the riboswitch, wherein the change in expression produces a signal.

11. The method of claim 10 wherein the signal is produced by a reporter protein expressed from the RNA linked to the riboswitch.

12. A method of inhibiting gene expression, the method comprising bringing into contact a compound and a cell, wherein the compound has the structure



wherein, when the compound is bound to a guanine-responsive riboswitch, R₇ serves as a hydrogen bond acceptor, R₁₀ serves as a hydrogen bond donor, R₁₁ serves as a hydrogen bond acceptor, R₁₂ serves as a hydrogen bond donor,

wherein R₁₃ is H, H₂ or is not present,

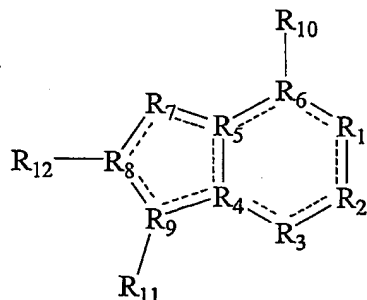
wherein R₁, R₂, R₃, R₄, R₅, R₆, R₈, and R₉ are each independently C, N, O, or S,

wherein ----- each independently represent a single or double bond,

wherein the compound is not guanine, hypoxanthine, or xanthine,

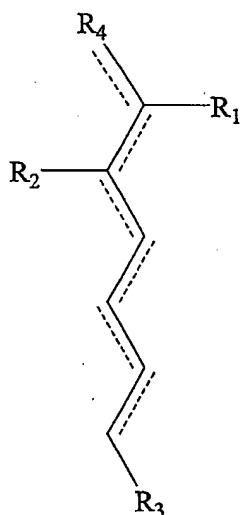
wherein the cell comprises a gene encoding an RNA comprising a guanine-responsive riboswitch, wherein the compound inhibits expression of the gene by binding to the guanine-responsive riboswitch.

13. A method of inhibiting gene expression, the method comprising bringing into contact a compound and a cell, wherein the compound has the structure



wherein, when the compound is bound to an adenine-responsive riboswitch, R₁, R₃ and R₇ serve as hydrogen bond acceptors, and R₁₀ and R₁₁ serve as hydrogen bond donors,
 wherein R₁₂ is H, H₂ or is not present,
 wherein R₁, R₂, R₃, R₄, R₅, R₆, R₈, and R₉ are each independently C, N, O, or S,
 wherein ----- each independently represent a single or double bond,
 wherein the compound is not adenine, 2,6-diaminopurine, or 2-amino purine,
 wherein the cell comprises a gene encoding an RNA comprising an adenine-responsive riboswitch, wherein the compound inhibits expression of the gene by binding to the adenine-responsive riboswitch.

14. A method of inhibiting gene expression, the method comprising
 bringing into contact a compound and a cell,
 wherein the compound has the structure



wherein R₂ and R₃ are each positively charged,
 wherein R₁ is negatively charged,
 wherein R₄ is C, N, O, or S,
 wherein ----- each independently represent a single or double bond,

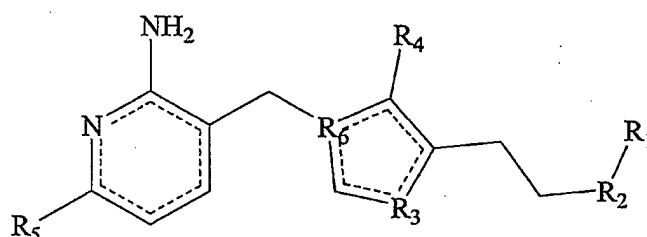
wherein the compound is not lysine,

wherein the cell comprises a gene encoding an RNA comprising a lysine-responsive riboswitch, wherein the compound inhibits expression of the gene by binding to the lysine-responsive riboswitch.

15. The method of claim 14 wherein R_2 and R_3 are each NH_3^+ and wherein R_1 is O^- .

16. A method of inhibiting gene expression, the method comprising bringing into contact a compound and a cell,

wherein the compound has the structure



wherein R_1 is positively charged,

wherein R_2 and R_3 are each independently C, O, or S,

wherein R_4 is CH_3 , NH_2 , OH, SH, H or not present,

wherein R_5 is CH_3 , NH_2 , OH, SH, or H,

wherein R_6 is C or N,

wherein ----- each independently represent a single or double bond,

wherein the compound is not TPP, TP or thiamine,

wherein the cell comprises a gene encoding an RNA comprising a thiamine pyrophosphate -responsive riboswitch, wherein the compound inhibits expression of the gene by binding to the thiamine pyrophosphate-responsive riboswitch.

17. The method of claim 16 wherein R_1 is phosphate, diphosphate or triphosphate.

18. A method comprising

(a) testing a compound for inhibition of gene expression of a gene encoding an RNA comprising a riboswitch, wherein the inhibition is via the riboswitch,

(b) inhibiting gene expression by bringing into contact a cell and a compound that inhibited gene expression in step (a),

wherein the cell comprises a gene encoding an RNA comprising a riboswitch, wherein the compound inhibits expression of the gene by binding to the riboswitch.

19. A method of identifying riboswitches, the method comprising

assess in-line spontaneous cleavage of an RNA molecule in the presence and absence of a compound, wherein the RNA molecule is encoded by a gene regulated by the compound, wherein a change in the pattern of in-line spontaneous cleavage of the RNA molecule indicates a riboswitch.

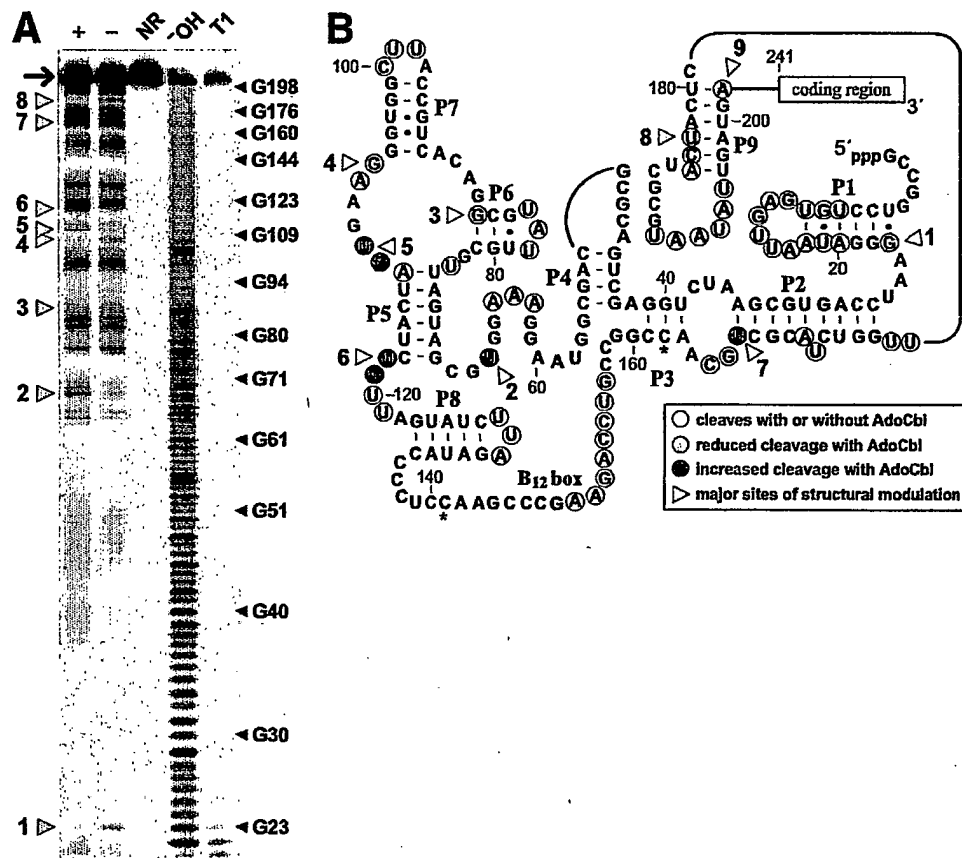
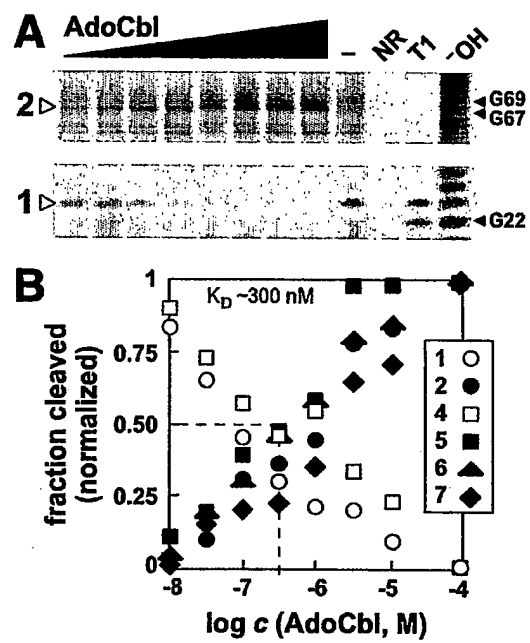
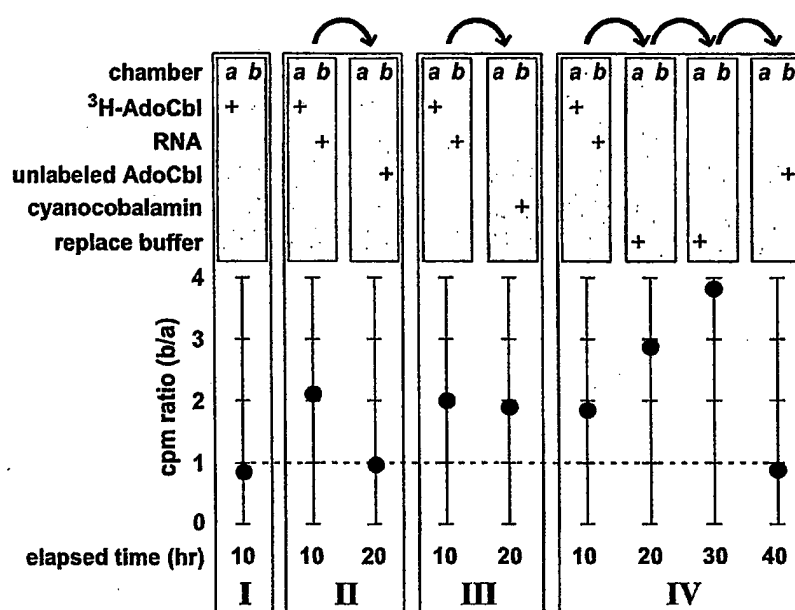


FIG. 1

**FIG. 2**

**FIG. 3**

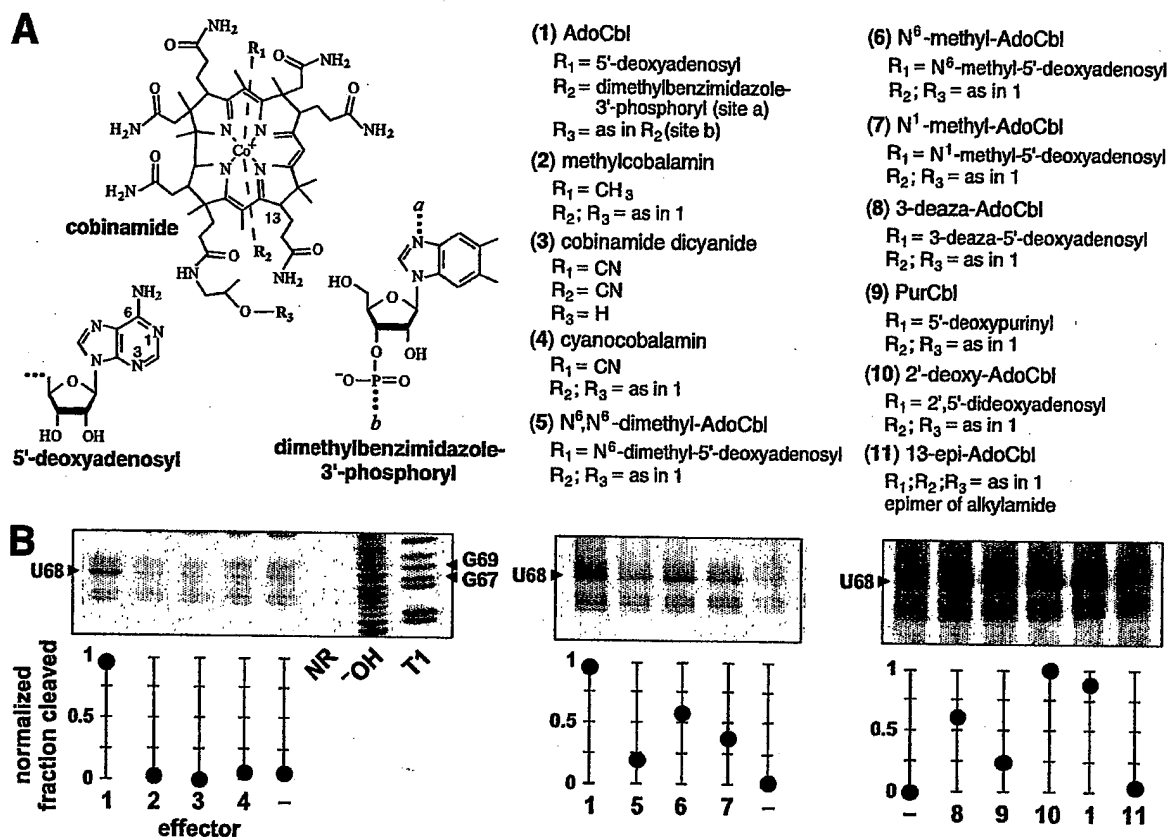


FIG. 4

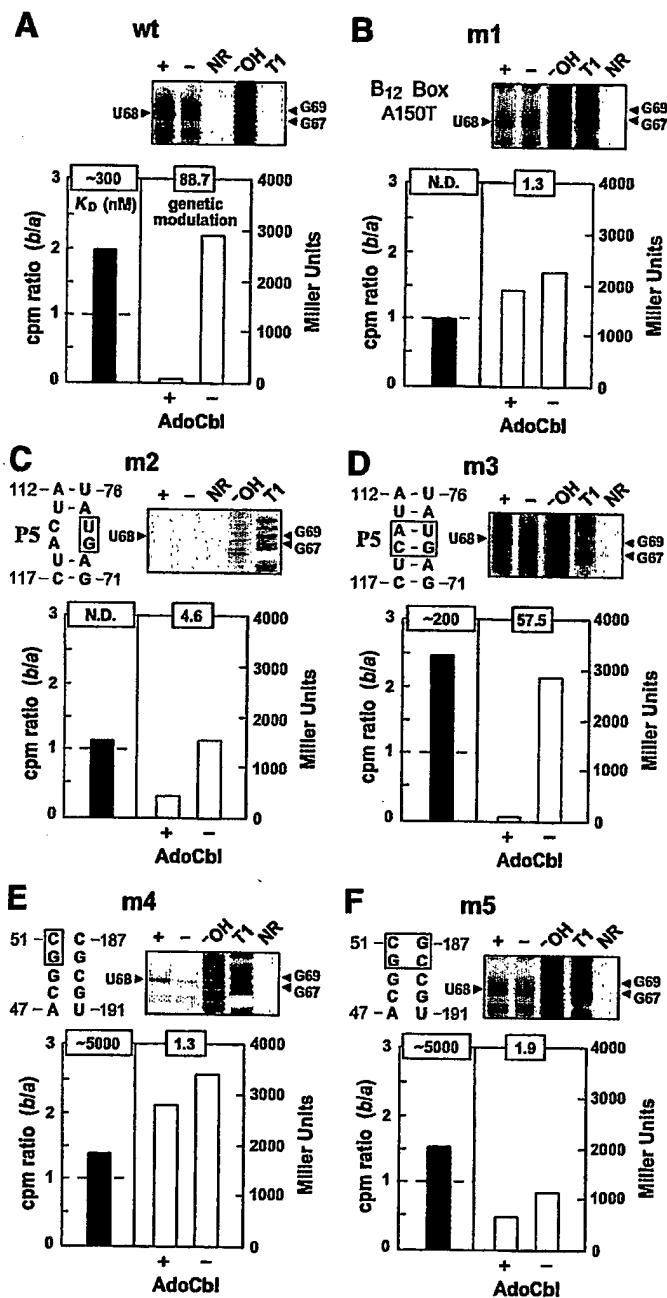


FIG. 5

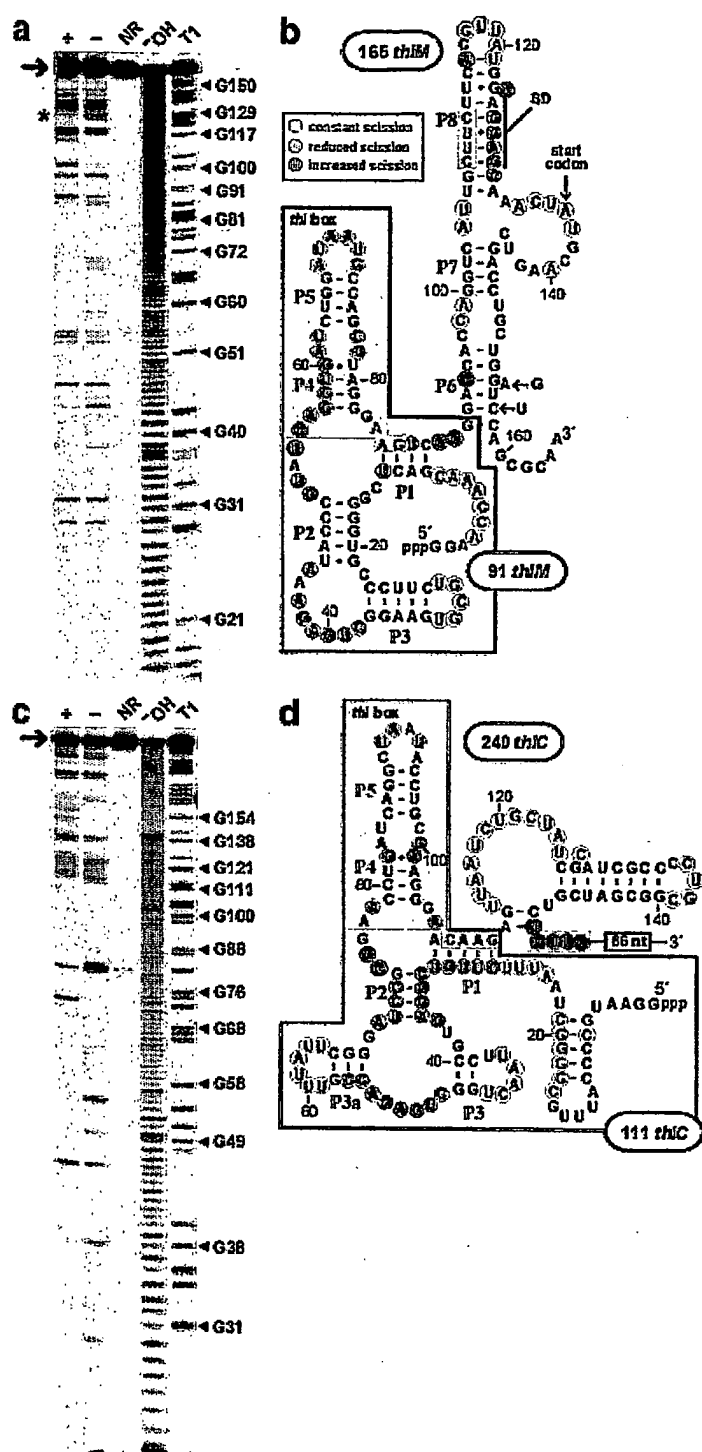


FIG. 6

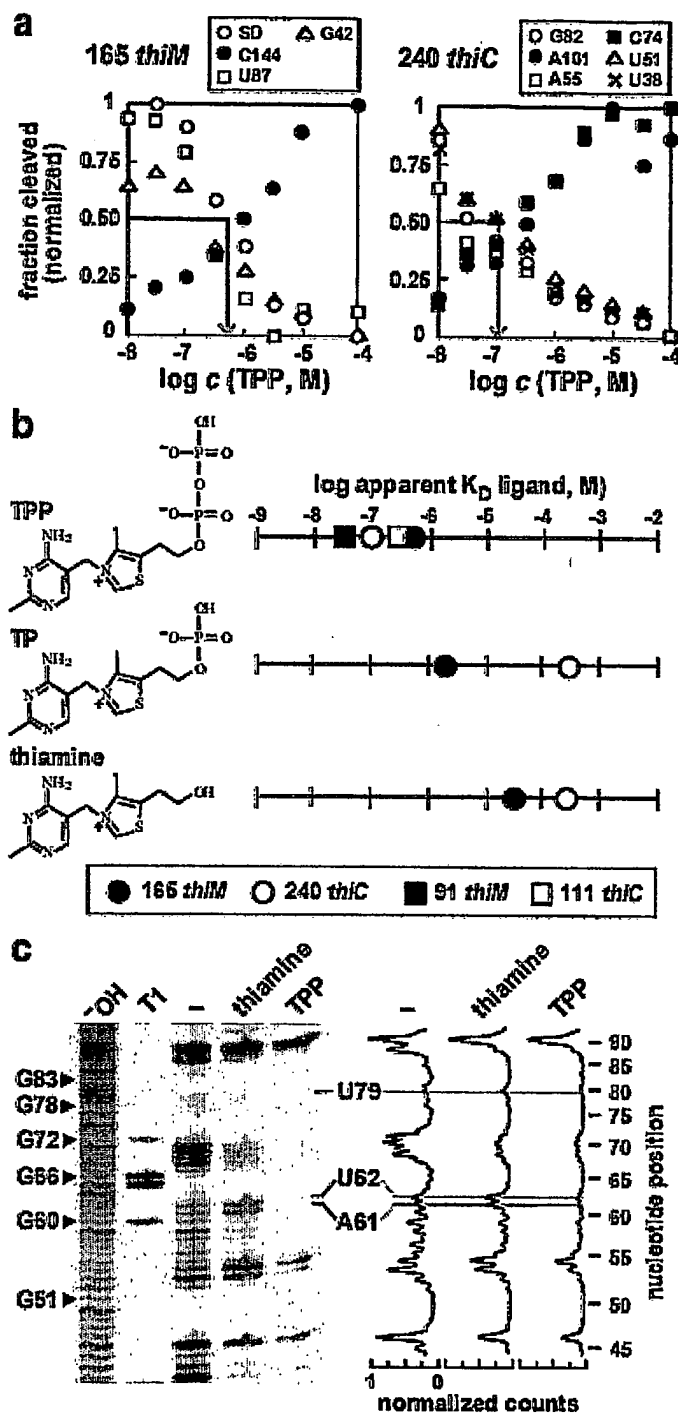


FIG. 7

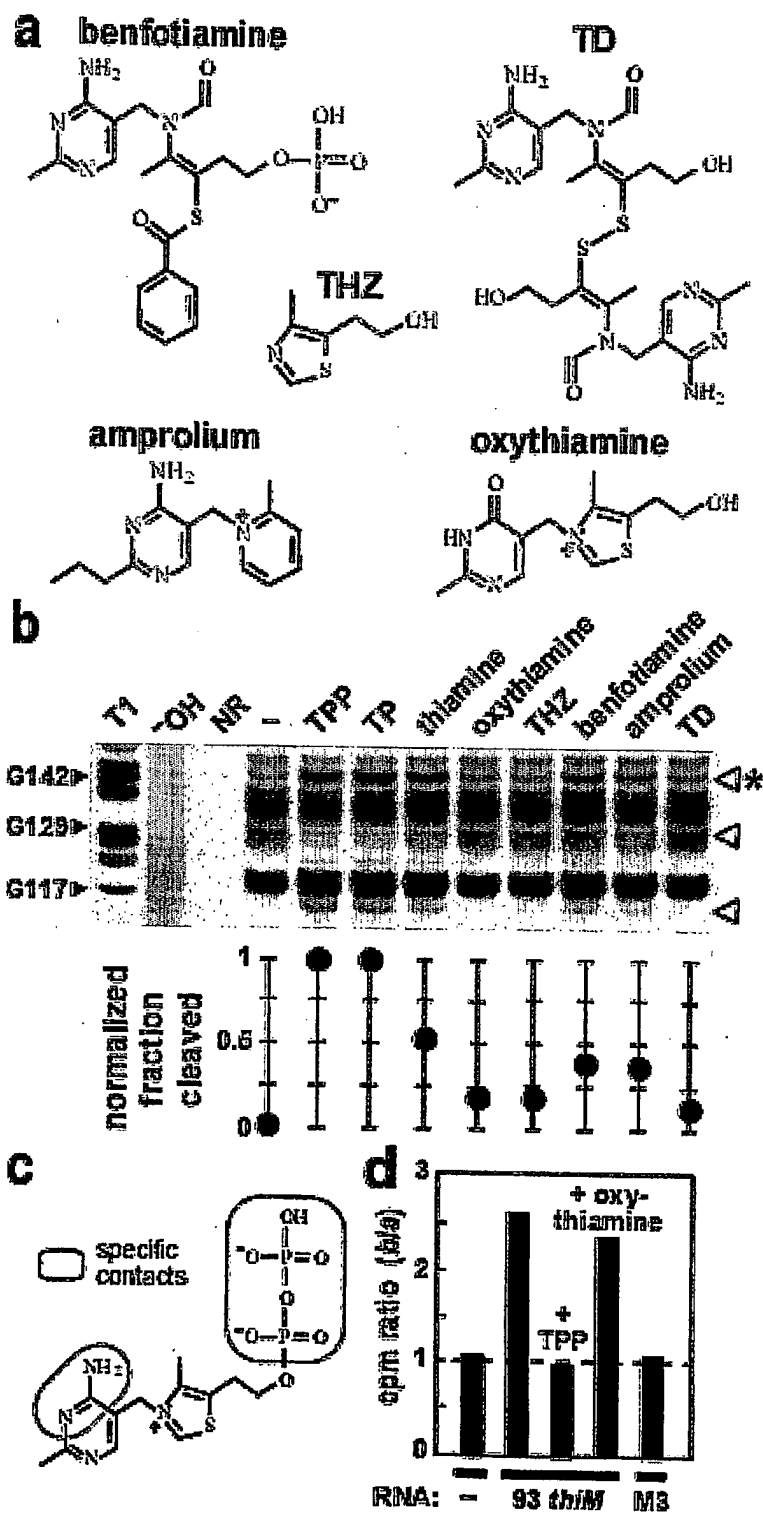


FIG. 8

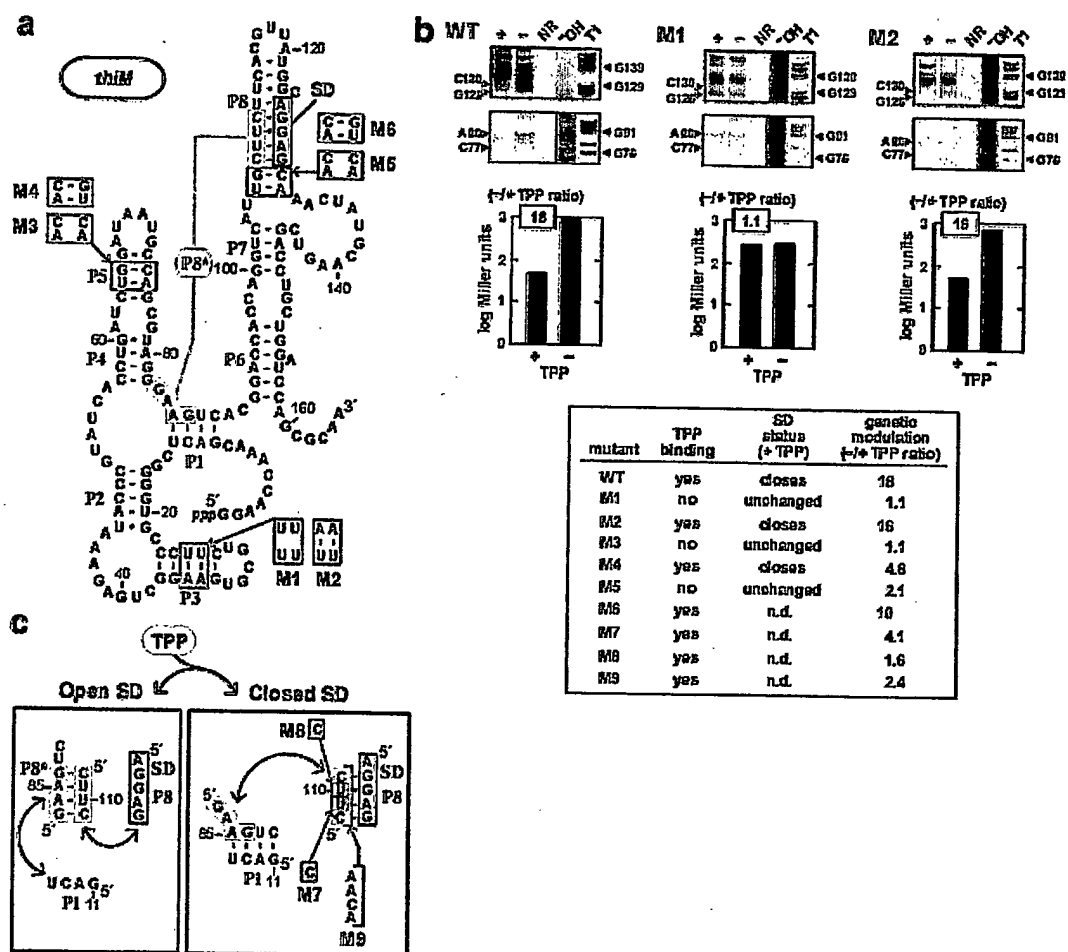


FIG. 9

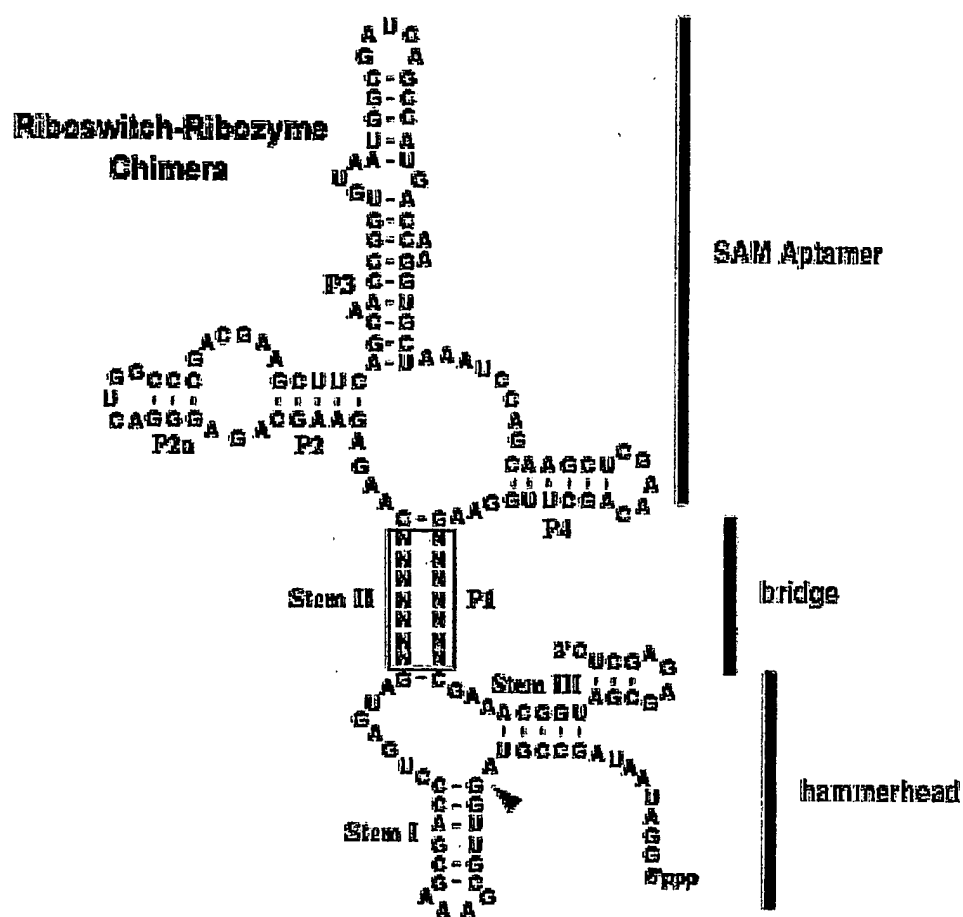


FIG. 10

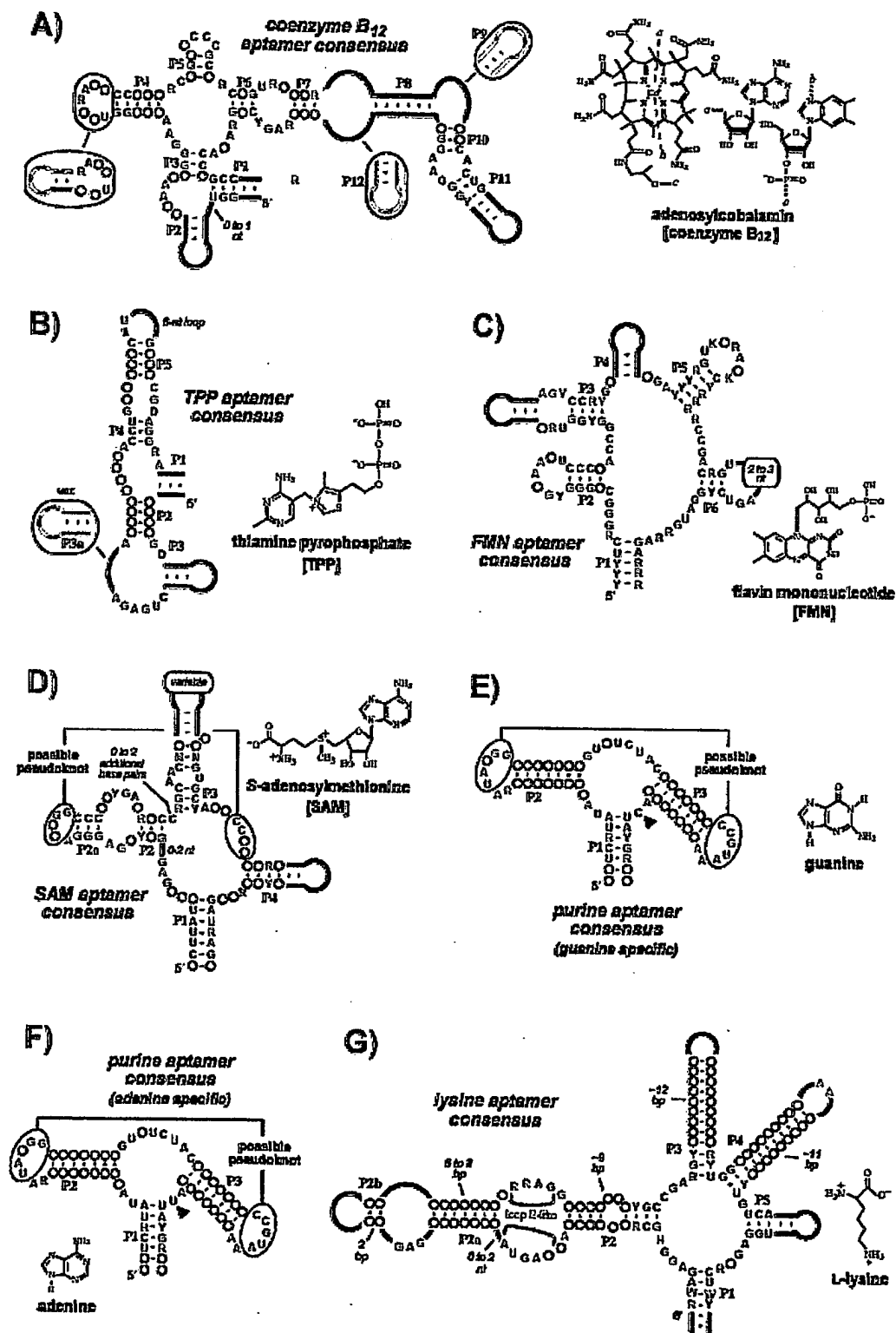


FIG. 11

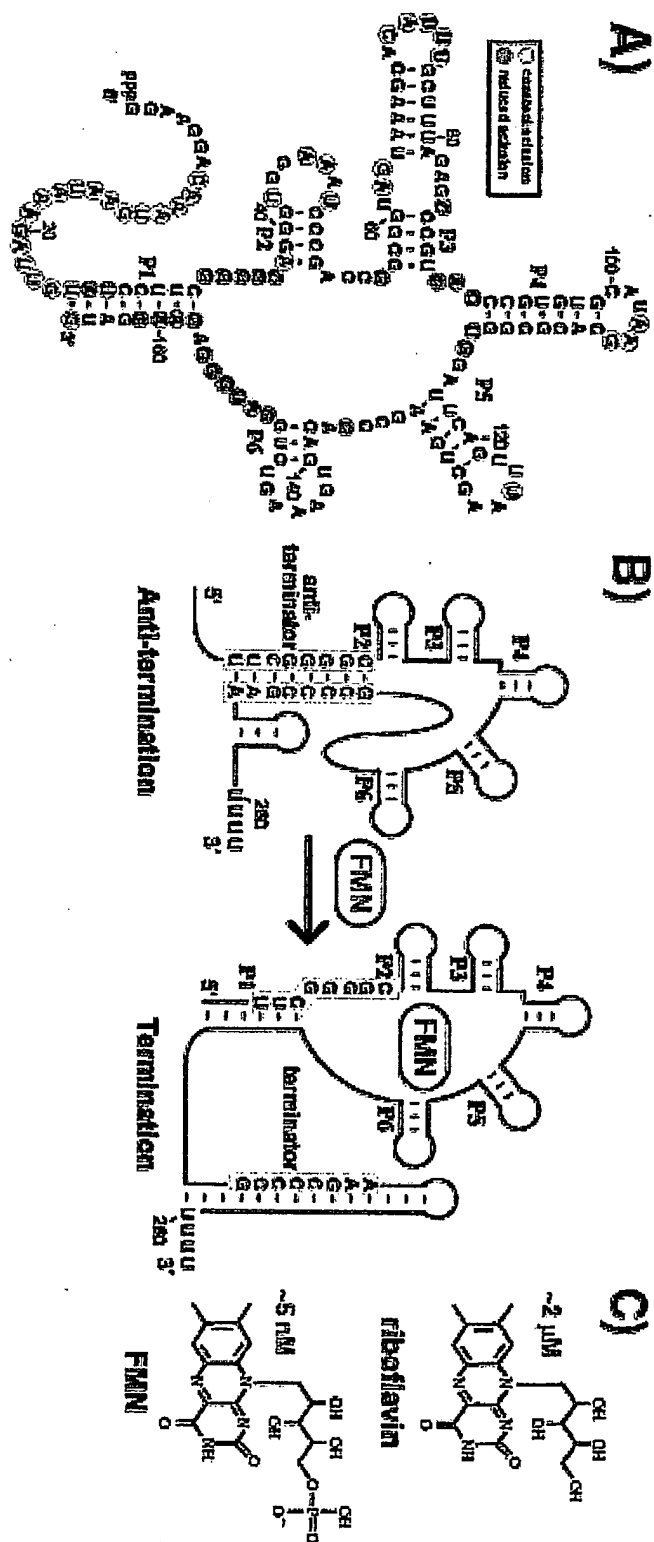


FIG. 12

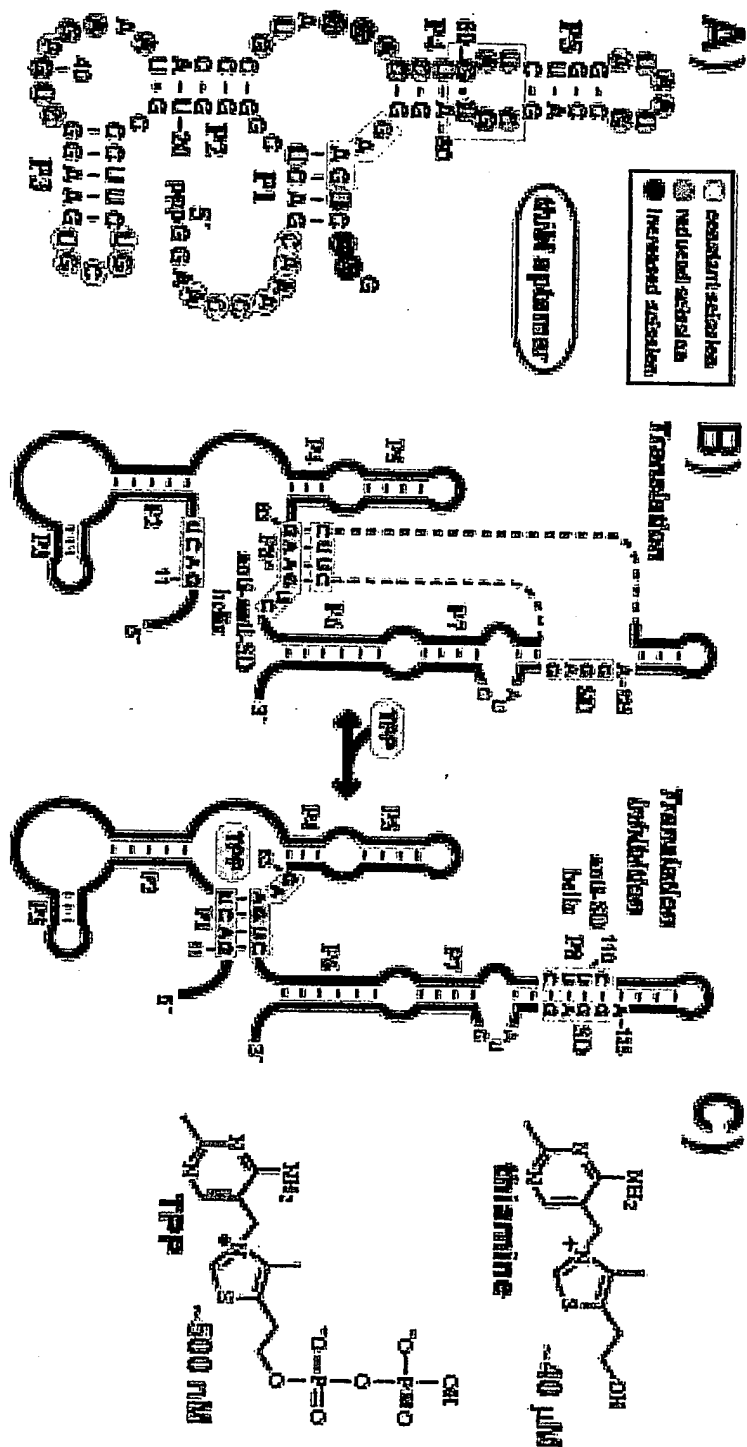


FIG. 13

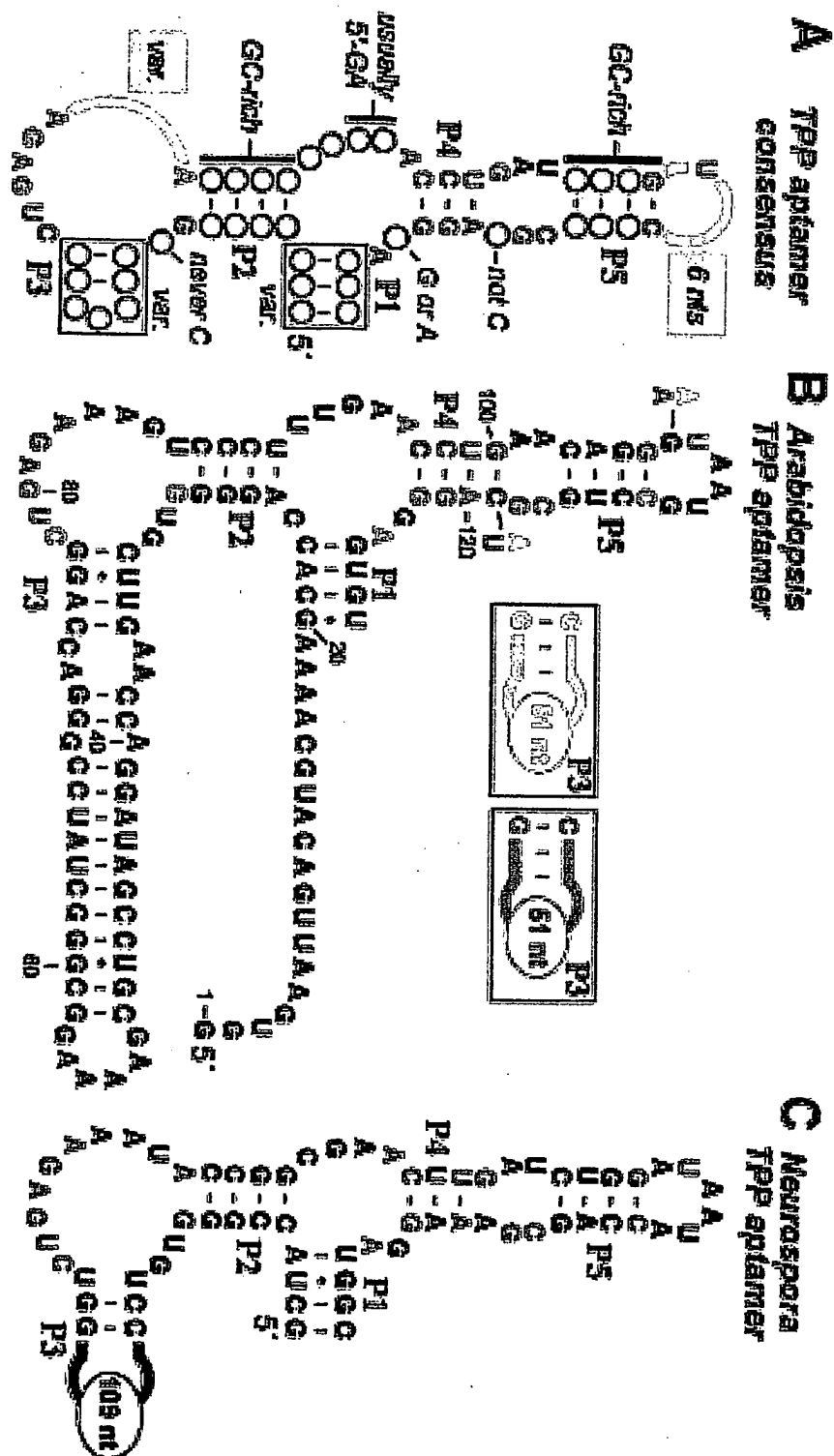


FIG. 14

ID	Position	Accession	Sequence Name	Gene	Location
Zco1	-	2189373	NC_000913.1	Escherichia coli K12 complete genome	thiM operon
Zco2	-	4193775	NC_000913.1	Escherichia coli K12 complete genome	thiC operon
Cac	-	3156135	NC_003030.1	Clostridium acetobutylicum ATCC824 complete genome	thiC
Ncr	+	664	AY007661.1	Neurospora crassa thiamine biosynthesis protein nmt-1 gene	nmt-1
Aor	+	622	AF217503.1	Aspergillus oryzae putative thiazole synthase (thiA) gene	thiA [thi4]
Fox	+	2156	AB033416.1	Fusarium oxysporum s135 gene for stress-responsive gene product	st135 [thi4]
Fzo	+	461	M39642.1	Fusarium solani ST135 protein gene	st135 [thi4]
Ath	-	78516	AC005496.3	Arabidopsis thaliana chromosome 2 clone T27A16 map m154	thiamine gene [thiC]
Pae	+	2296	AF264021.1	Poa secunda putative thiamine biosynthesis protein ThiO mRNA	putative thiC
Oza	-	91318	AC084406.7	Oryza sativa chromosome 3 BAC OSJNBa0067E01 genomic sequence	putative thiC

P3

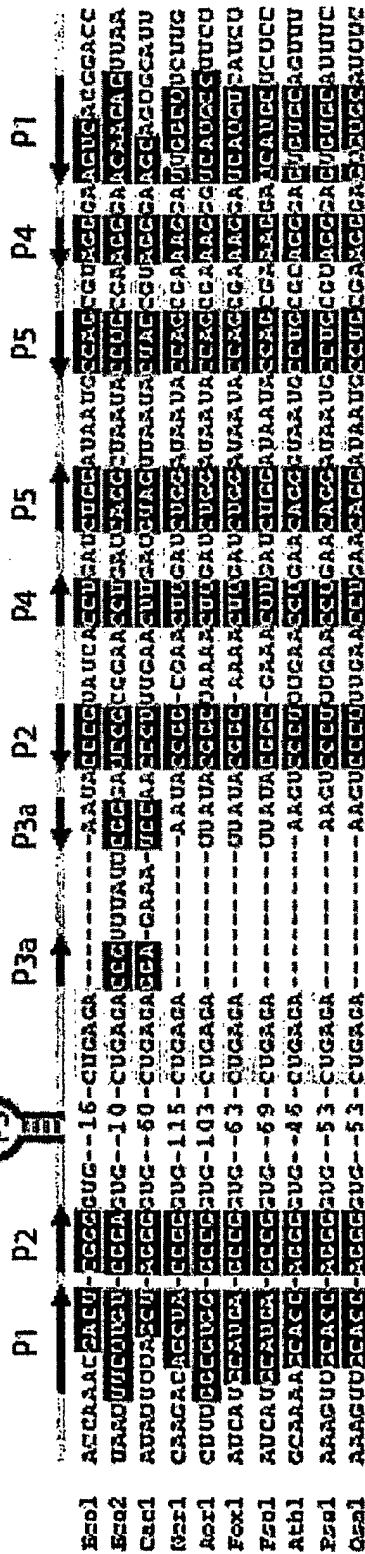


FIG. 15

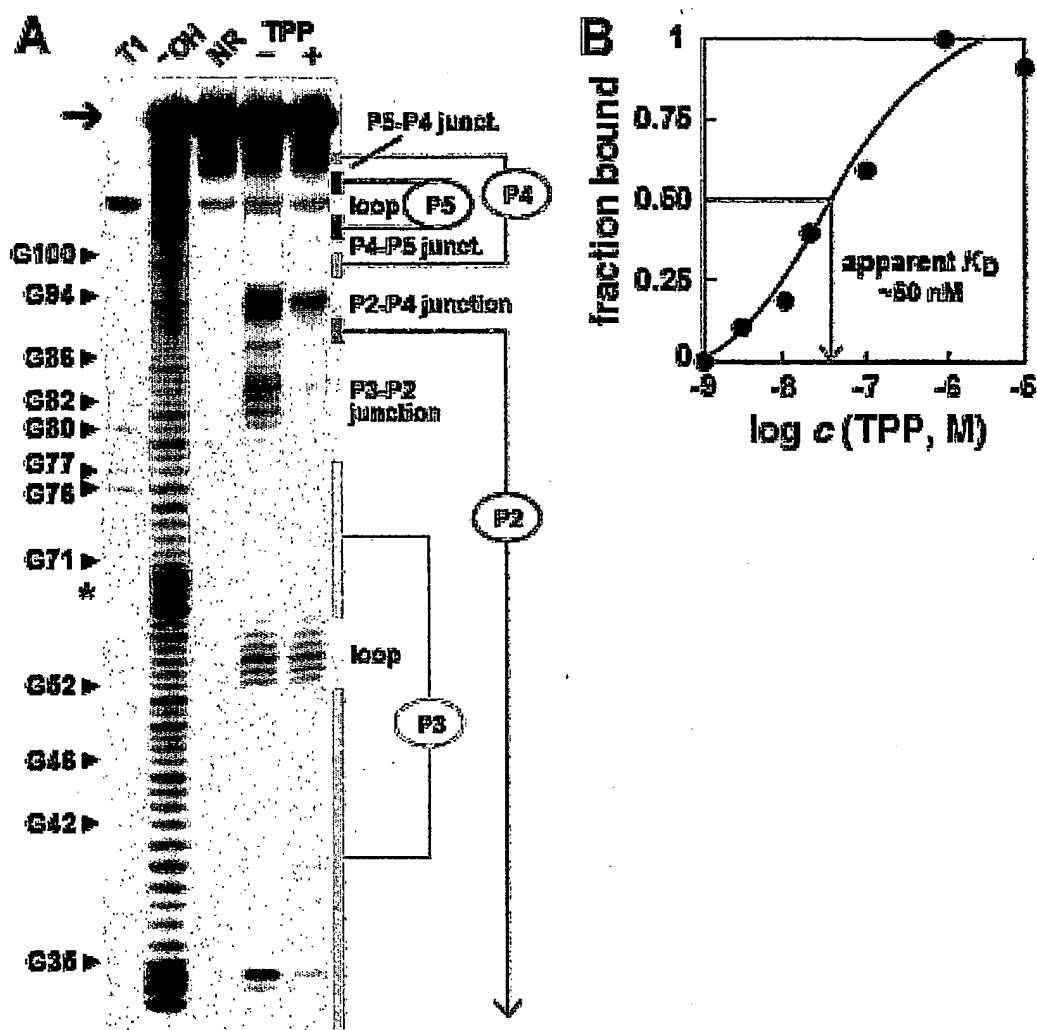


FIG. 16

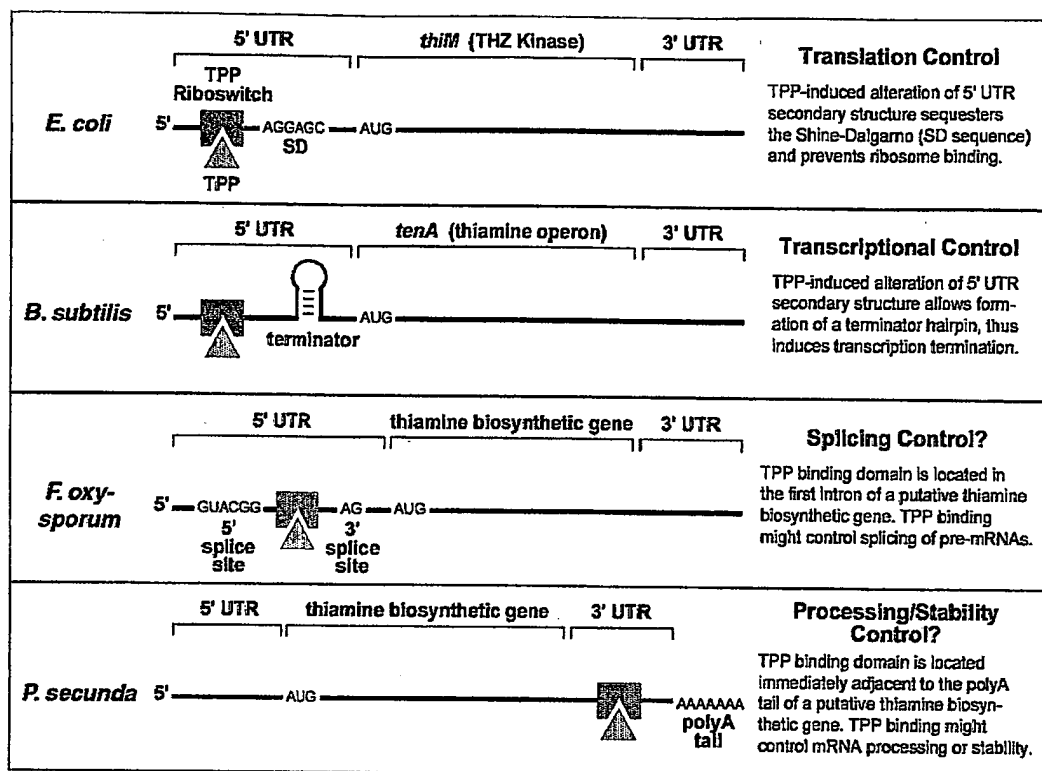


FIG. 17

1	BA 0845	CGGUGAGGU	AGGCU	CTUCAUU	AAG-AGUA	UCAAUUCG	G-AG	AGAGUG	ATGAUGA	AGGAUUA	GAAGGAGUAT	UCCGA	
2	BA 1ysA	CUCAAAGGU	AGSCC	UUGG	AAAGAGUA	ACCUAUGG	G-AG	UUAAGG	AA	GGUU	GAAGGGAUUA	UCCGA	
3	BA 1ysP	CGUAGGAGU	AGGCU	CAUUDUU	AAG-AGUA	AAAGGAGC	G-AGAG	AGGAGAA	UCCUAGA	UCCGCUU	GAAGGGAUUA	UCCGA	
4	BH dapA	AGUGAGAU	AGGU	AAAC	AAG-AGUA	CACAUAU	GG-AG	GAGAGGAGA	UCCUGAG	AAUUG	GAAGGGAUUA	UCCGA	
5	BH 1ysC	AGUGAGAU	AGGU	AAAC	AAG-AGUA	ACAGUC	UGAG-AG	AAUAGAGAUU	UUGAC	GAAGG	GAAGGGAUUA	UCCGA	
6	BH nhaC	AGAGGAGU	AGGA	CGUUUU	AAG-AGUA	CGCCUUG	G-AGG	AGGACAA	CCAGGA	UAGGAG	CGAAGGGAUUA	UCCGA	
7	BS 1ysC	GGUGAGAU	AGGU	CAUUC	AAG-AGUA	CGCUUU	GG-AG	AAAGAGUA	UUCUGGAA	AAAGG	UGAAGGGAUUA	UCCGA	
8	CA 1ysA	AGUGAGAU	AGGU	UUAAGU	AAG-AGUA	CGCUUU	G-AG	UUGGCA	AAUUGA	UAGGAG	UAAAGGGAUUA	UCCGA	
9	CP 1ysA	GAGGAGAU	AGGU	CGUAUU	AAG-AGUA	GUCAUA	AGUG	UGGCA	AG	9	GAAGG	GAAGGGAUUA	UCCGA
10	CP 1ysP	AACUGAGAU	AGGC	GAUUAU	AAG-AGUA	UCUUUG	AG-AG	GUAGCA	AUUGAA	GAAGG	UGAAGGGAUUA	UCCGA	
11	CP nhaC	AAAGAGAU	AGGC	GAGAUU	AAG-AGUA	CUAUAU	GG-AG	UUAAGU	AGCGUAGAG	UUAAG	GAAGGGAUUA	UCCGA	
12	EC 1ysC	CAGGAGAU	AGGC	UUGCCC	A	AGUAGC	GGUUG	G-AG	GGGAG	UCCUGA	GAAGGGAUUA	UCCGA	
13	HI nhaC	UACAGAGU	AGGC	UUAUUU	AUA-AGUA	UUUUUU	AG-AG	UGGAAU	CGAAGA	GAAGG	UGGAGGGAUUA	UCCGA	
14	OI dapA	GUUUUGAU	AGGU	GGGAGC	AUC-AGUA	UAUAGC	GGAG	GGAAUGAG	CG	8	GAAGG	GAAGGGAUUA	UCCGA
15	OI nhaC	UCCGAGAU	AGGA	UUCAGC	AUC-AGUA	UUGUUU	AG-AG	UGGAAU	CGAAGA	GAAGG	UGGAGGGAUUA	UCCGA	
16	PH nhaC	UACUUGAU	AGGA	UUCAGC	AUC-AGUA	UUGUUU	G-AG	UGGAAU	CGAAGA	GAAGG	UGGAGGGAUUA	UCCGA	
17	SA 1ysC	AUAGAGAU	AGGC	UUCAGC	AUC-AGUA	UUGUUU	G-AG	UGGAAU	CGAAGA	GAAGG	UGGAGGGAUUA	UCCGA	
18	SA 1ysP	AUAGAGAU	AGGC	UUCAGC	AUC-AGUA	UUGUUU	G-AG	UGGAAU	CGAAGA	GAAGG	UGGAGGGAUUA	UCCGA	
19	SE 1ysC	AGAGAGAU	AGGU	UUCAGC	AUC-AGUA	UUGUUU	G-AG	UGGAAU	CGAAGA	GAAGG	UGGAGGGAUUA	UCCGA	
20	SE 1ysP	AGAGAGAU	AGGU	UUCAGC	AUC-AGUA	UUGUUU	G-AG	UGGAAU	CGAAGA	GAAGG	UGGAGGGAUUA	UCCGA	
21	SP 1ysC	CAGGAGAU	AGGC	UUGCCC	A	AGUAGC	GGUUG	G-AG	GGGAG	UCCUGA	GAAGGGAUUA	UCCGA	
22	SO 1ysC	CAGGAGAU	AGGC	UUGCCC	A	AGUAGC	GGUUG	G-AG	GGGAG	UCCUGA	GAAGGGAUUA	UCCGA	
23	SO nhaC	CGUAGAGU	AGGC	UUGCCC	A	AGUAGC	GGUUG	G-AG	GGGAG	UCCUGA	GAAGGGAUUA	UCCGA	
24	TM asd	UCCGAGAU	AGGC	UUGCCC	A	AGUAGC	GGUUG	G-AG	GGGAG	UCCUGA	GAAGGGAUUA	UCCGA	
25	TT 1ysA	AGGAGAGU	AGGC	UUGCCC	A	AGUAGC	GGUUG	G-AG	GGGAG	UCCUGA	GAAGGGAUUA	UCCGA	
26	TT pspF	CGGAGAGU	AGGA	UUGCCC	A	AGUAGC	GGUUG	G-AG	GGGAG	UCCUGA	GAAGGGAUUA	UCCGA	
27	VC 1ysC	UUCAGAGU	AGGA	UUGCCC	A	AGUAGC	GGUUG	G-AG	GGGAG	UCCUGA	GAAGGGAUUA	UCCGA	
28	VC nhaC	UUCGAGAGU	AGGA	UUGCCC	A	AGUAGC	GGUUG	G-AG	GGGAG	UCCUGA	GAAGGGAUUA	UCCGA	
29	VC nhaC	CUUUGAGAGU	AGGA	UUGCCC	A	AGUAGC	GGUUG	G-AG	GGGAG	UCCUGA	GAAGGGAUUA	UCCGA	
30	VV 1ysC	UUCUGAGAGU	AGGA	UUGCCC	A	AGUAGC	GGUUG	G-AG	GGGAG	UCCUGA	GAAGGGAUUA	UCCGA	
31	VV nhaC	UUGCAGAGU	AGGC	UUGGUA	AAG-AGUA	ACUUAUUU	G-GC	UGAU	GGCAUG	GAUUAU	GAAGGGAUUA	UCCGA	

1	P3	P3'	P4	P4'	P5	P5'	P1'
1	AGUAGAGU	CCACAU	AGUAGAGU	UUAUAU	AGUAGAGU	UUAUAU	AGUAGAGU
2	AGUAGAGU	CCACAU	AGUAGAGU	UUAUAU	AGUAGAGU	UUAUAU	AGUAGAGU
3	AGUAGAGU	CCACAU	AGUAGAGU	UUAUAU	AGUAGAGU	UUAUAU	AGUAGAGU
4	AGUAGAGU	CCACAU	AGUAGAGU	UUAUAU	AGUAGAGU	UUAUAU	AGUAGAGU
5	AGUAGAGU	CCACAU	AGUAGAGU	UUAUAU	AGUAGAGU	UUAUAU	AGUAGAGU
6	AGUAGAGU	CCACAU	AGUAGAGU	UUAUAU	AGUAGAGU	UUAUAU	AGUAGAGU
7	AGUAGAGU	CCACAU	AGUAGAGU	UUAUAU	AGUAGAGU	UUAUAU	AGUAGAGU
8	AGUAGAGU	CCACAU	AGUAGAGU	UUAUAU	AGUAGAGU	UUAUAU	AGUAGAGU
9	AGUAGAGU	CCACAU	AGUAGAGU	UUAUAU	AGUAGAGU	UUAUAU	AGUAGAGU
10	AGUAGAGU	CCACAU	AGUAGAGU	UUAUAU	AGUAGAGU	UUAUAU	AGUAGAGU
11	AGUAGAGU	CCACAU	AGUAGAGU	UUAUAU	AGUAGAGU	UUAUAU	AGUAGAGU
12	AGUAGAGU	CCACAU	AGUAGAGU	UUAUAU	AGUAGAGU	UUAUAU	AGUAGAGU
13	AGUAGAGU	CCACAU	AGUAGAGU	UUAUAU	AGUAGAGU	UUAUAU	AGUAGAGU
14	AGUAGAGU	CCACAU	AGUAGAGU	UUAUAU	AGUAGAGU	UUAUAU	AGUAGAGU
15	AGUAGAGU	CCACAU	AGUAGAGU	UUAUAU	AGUAGAGU	UUAUAU	AGUAGAGU
16	AGUAGAGU	CCACAU	AGUAGAGU	UUAUAU	AGUAGAGU	UUAUAU	AGUAGAGU
17	AGUAGAGU	CCACAU	AGUAGAGU	UUAUAU	AGUAGAGU	UUAUAU	AGUAGAGU
18	AGUAGAGU	CCACAU	AGUAGAGU	UUAUAU	AGUAGAGU	UUAUAU	AGUAGAGU
19	AGUAGAGU	CCACAU	AGUAGAGU	UUAUAU	AGUAGAGU	UUAUAU	AGUAGAGU
20	AGUAGAGU	CCACAU	AGUAGAGU	UUAUAU	AGUAGAGU	UUAUAU	AGUAGAGU
21	AGUAGAGU	CCACAU	AGUAGAGU	UUAUAU	AGUAGAGU	UUAUAU	AGUAGAGU
22	AGUAGAGU	CCACAU	AGUAGAGU	UUAUAU	AGUAGAGU	UUAUAU	AGUAGAGU
23	AGUAGAGU	CCACAU	AGUAGAGU	UUAUAU	AGUAGAGU	UUAUAU	AGUAGAGU
24	AGUAGAGU	CCACAU	AGUAGAGU	UUAUAU	AGUAGAGU	UUAUAU	AGUAGAGU
25	AGUAGAGU	CCACAU	AGUAGAGU	UUAUAU	AGUAGAGU	UUAUAU	AGUAGAGU
26	AGUAGAGU	CCACAU	AGUAGAGU	UUAUAU	AGUAGAGU	UUAUAU	AGUAGAGU
27	AGUAGAGU	CCACAU	AGUAGAGU	UUAUAU	AGUAGAGU	UUAUAU	AGUAGAGU
28	AGUAGAGU	CCACAU	AGUAGAGU	UUAUAU	AGUAGAGU	UUAUAU	AGUAGAGU
29	AGUAGAGU	CCACAU	AGUAGAGU	UUAUAU	AGUAGAGU	UUAUAU	AGUAGAGU
30	AGUAGAGU	CCACAU	AGUAGAGU	UUAUAU	AGUAGAGU	UUAUAU	AGUAGAGU
31	AGUAGAGU	CCACAU	AGUAGAGU	UUAUAU	AGUAGAGU	UUAUAU	AGUAGAGU

FIG. 18

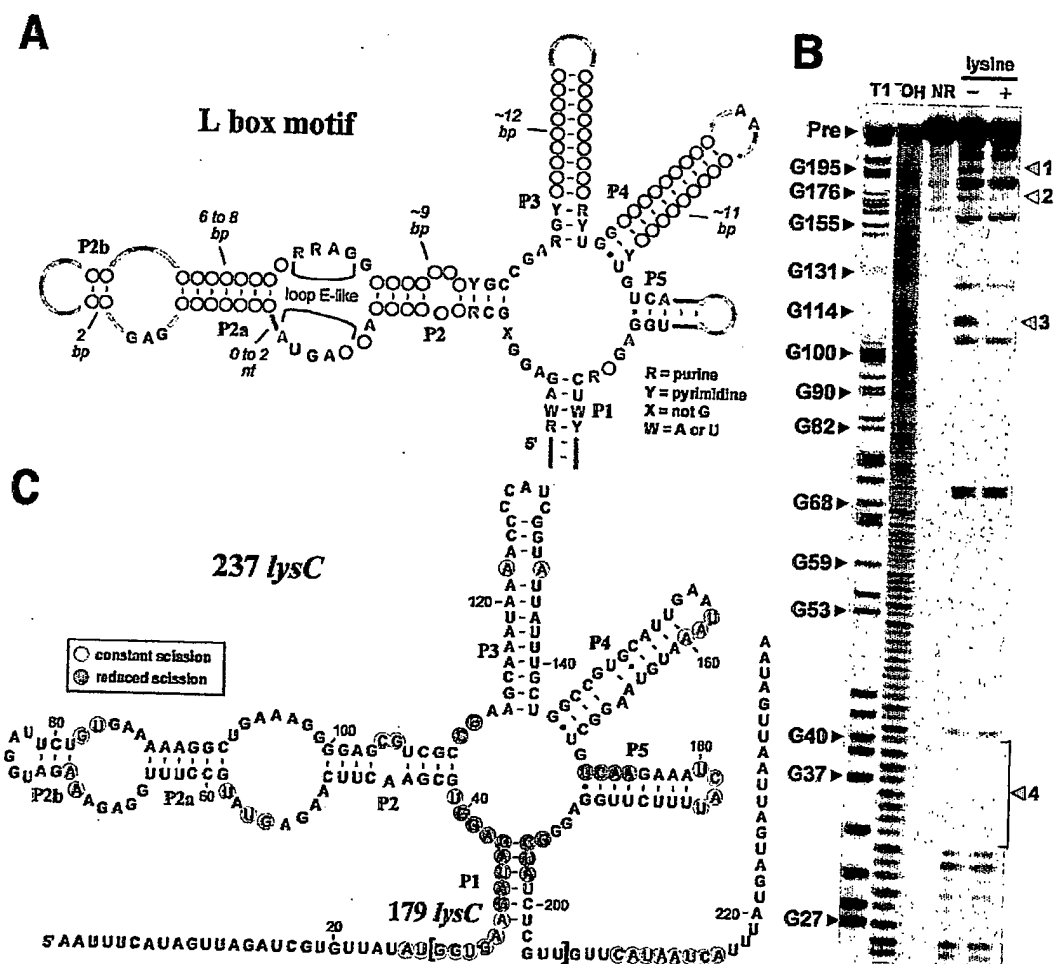


FIG. 19

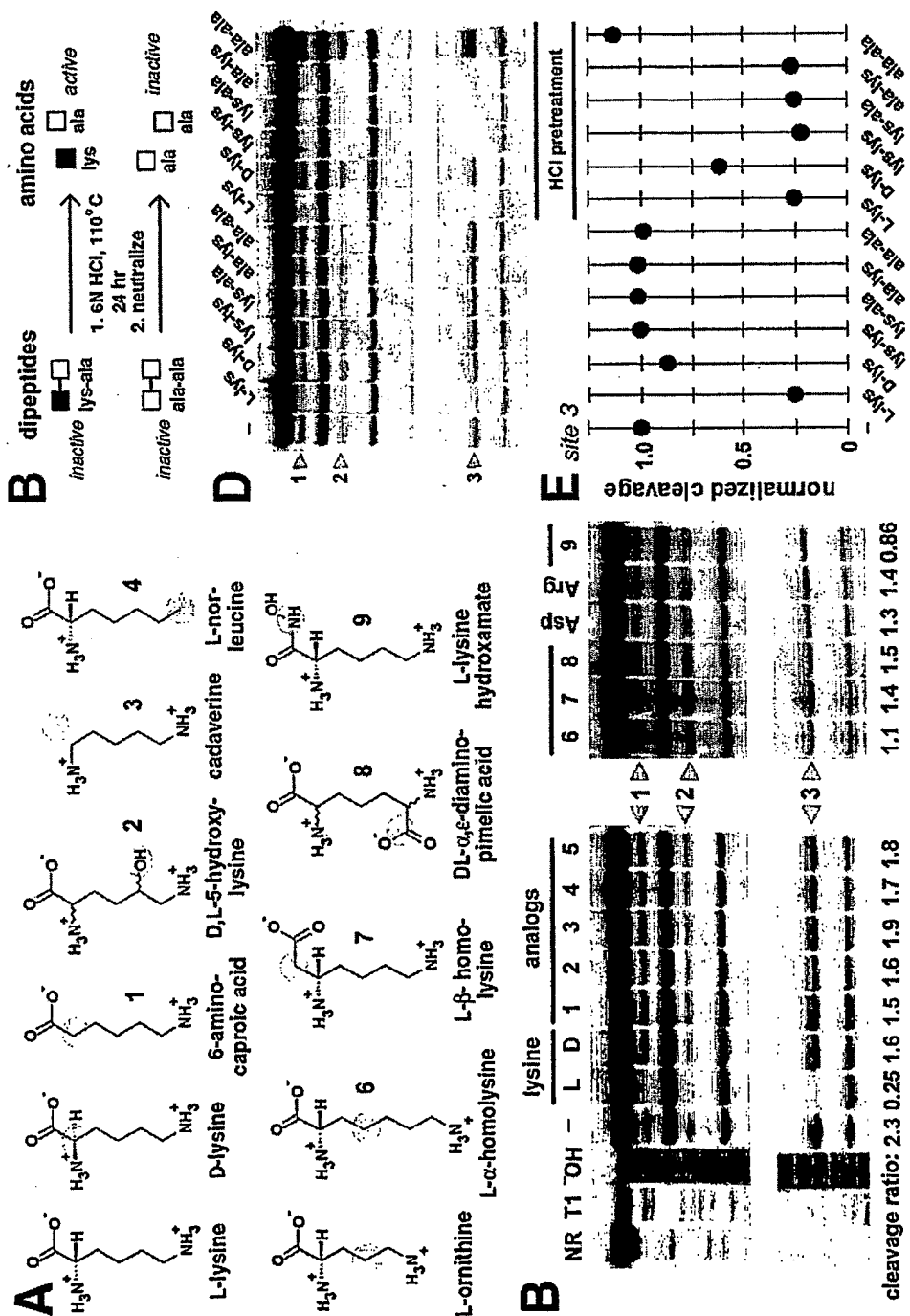
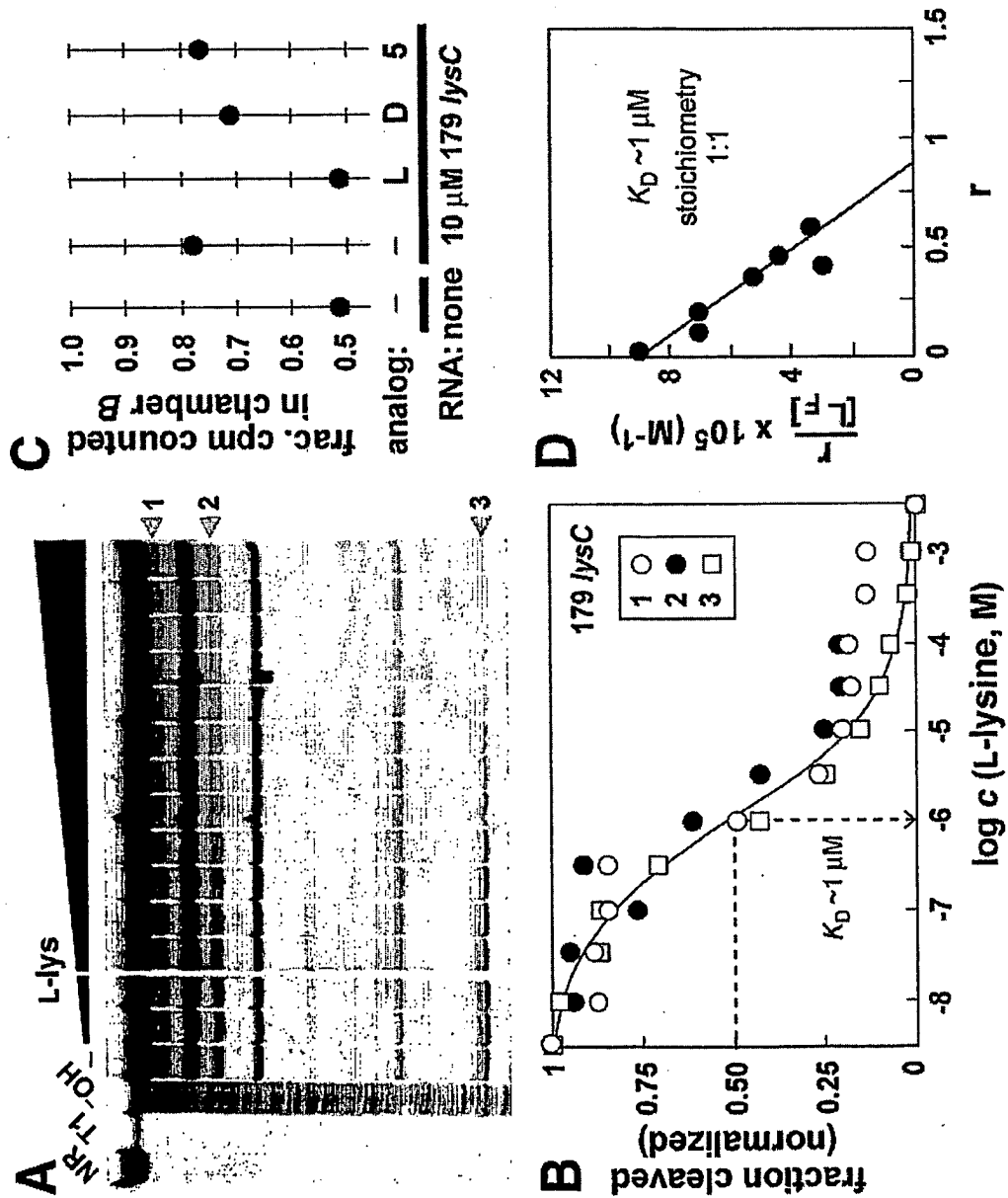


FIG. 20



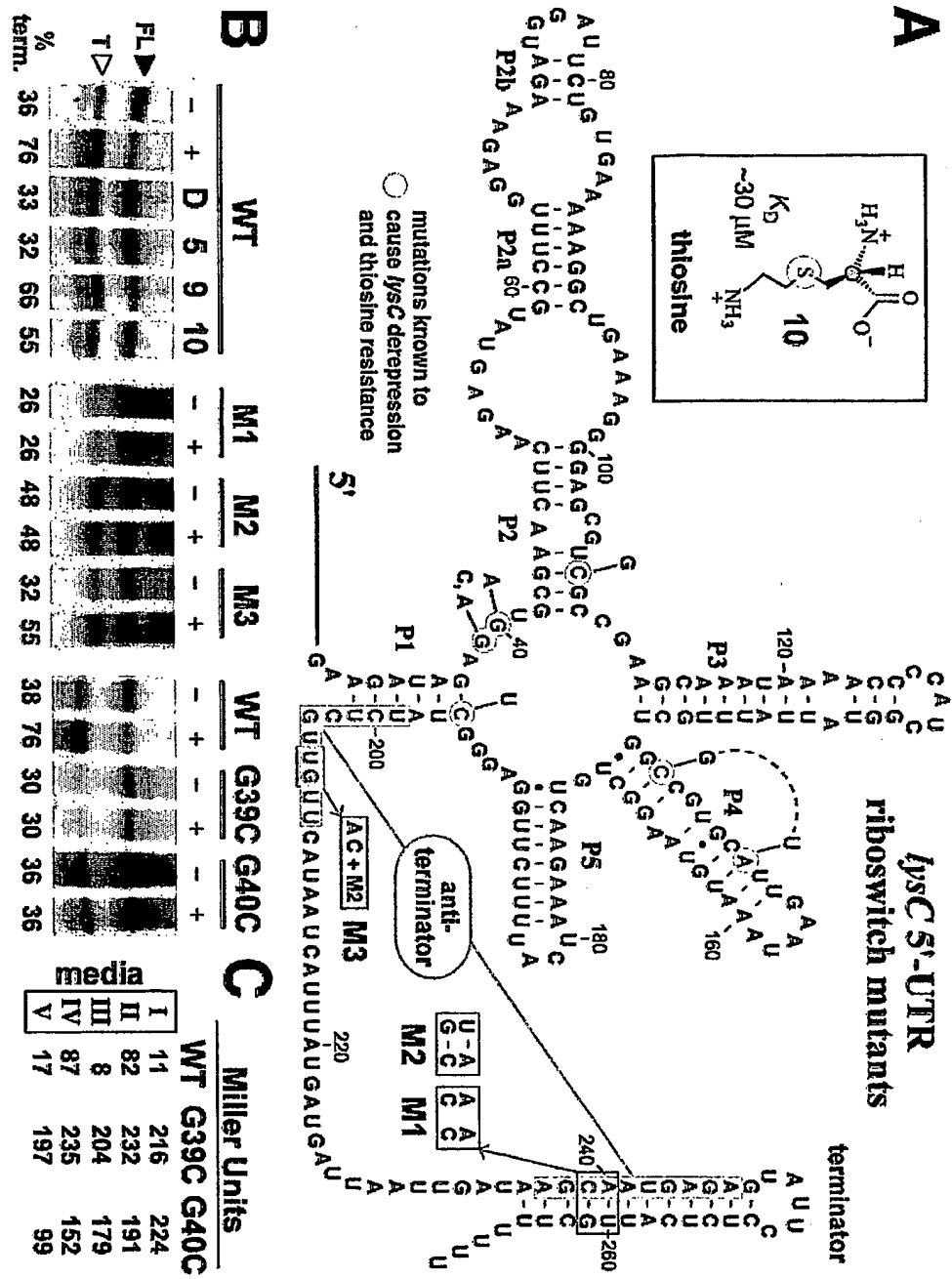


FIG. 22

FIG. 23

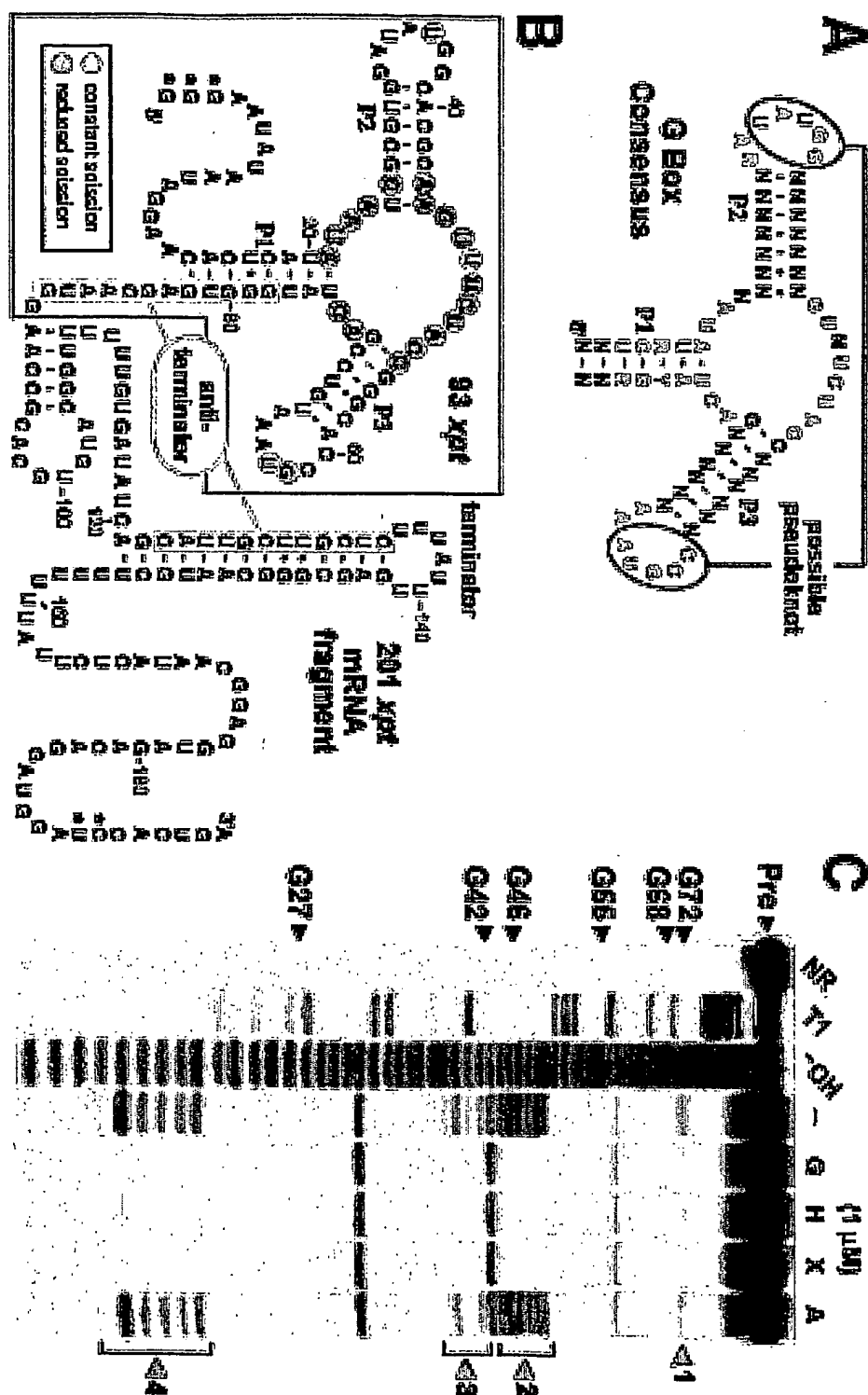


FIG. 24

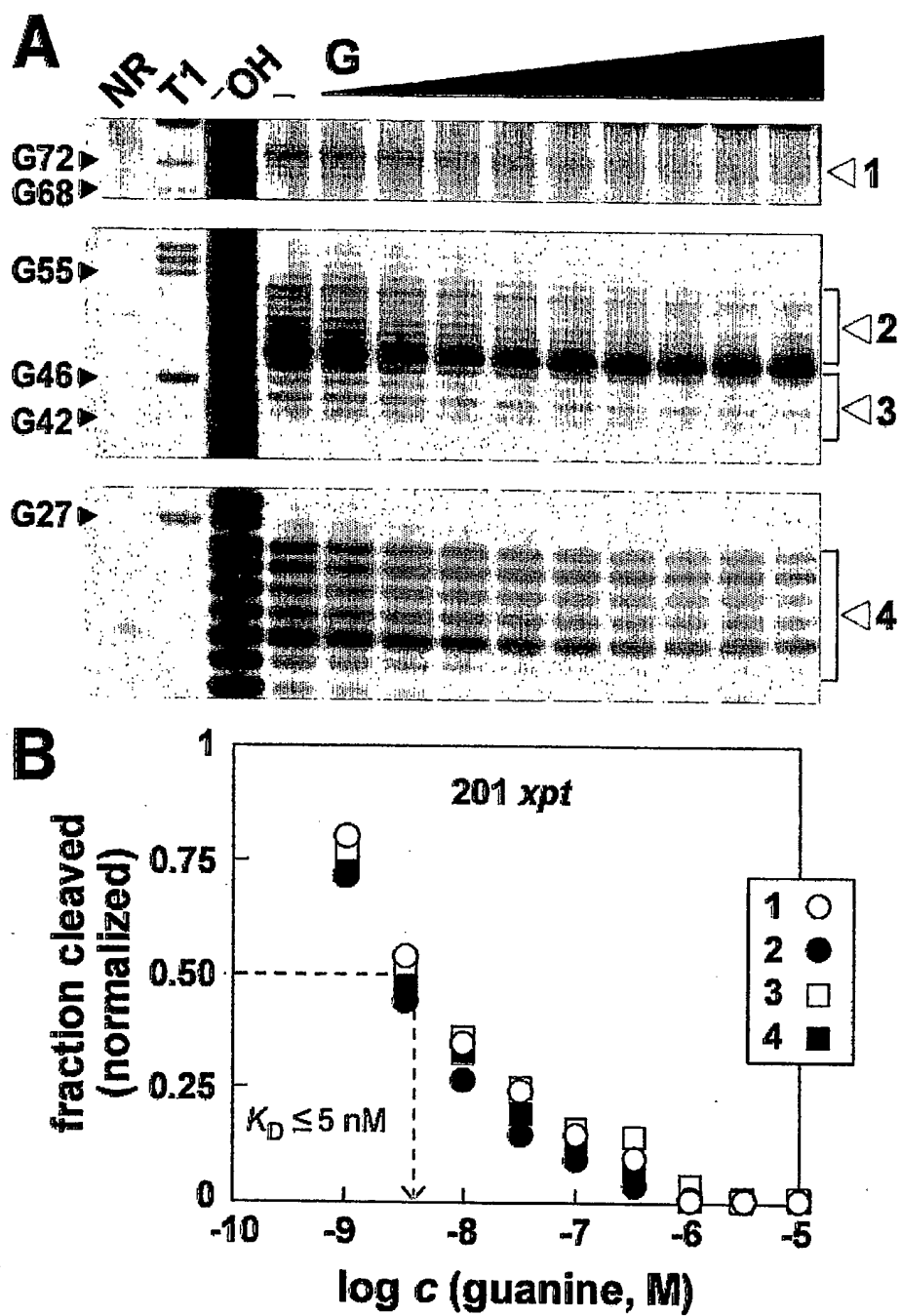


FIG. 25

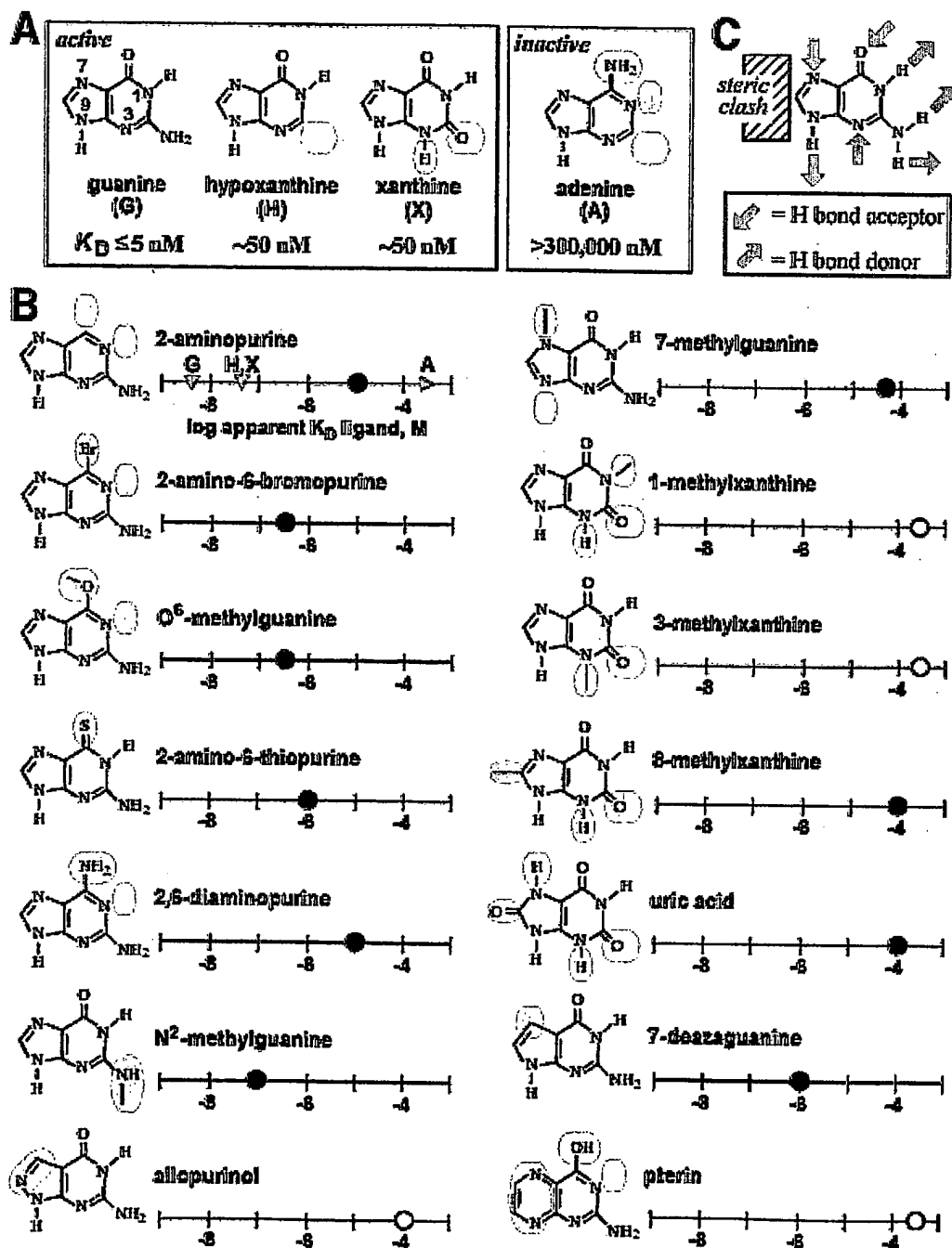


FIG. 26

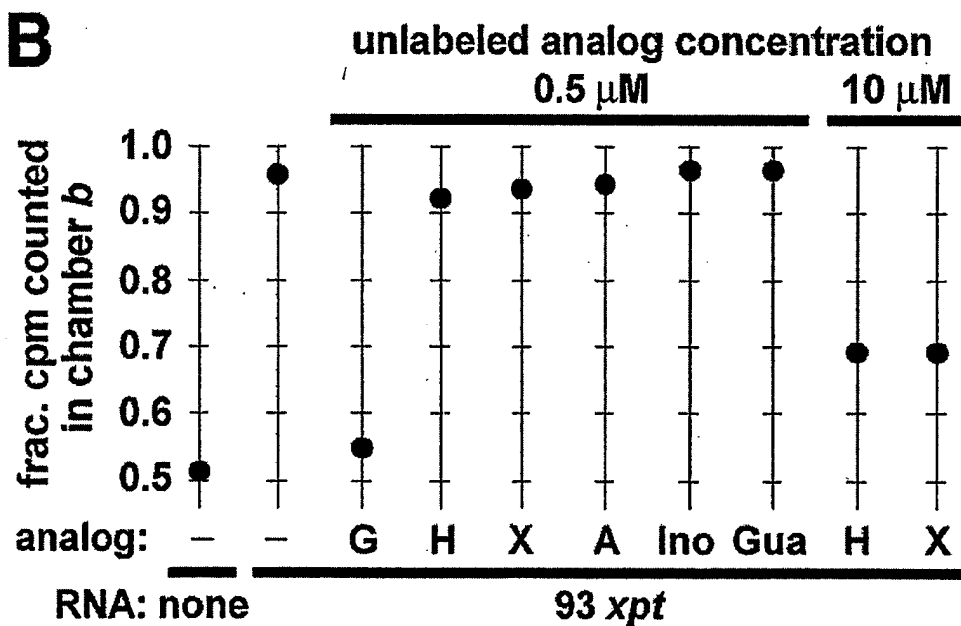
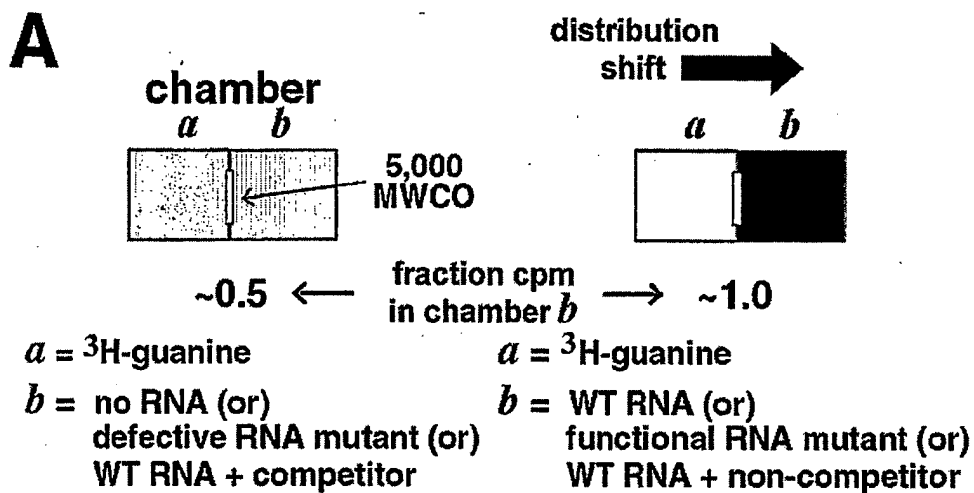


FIG. 27

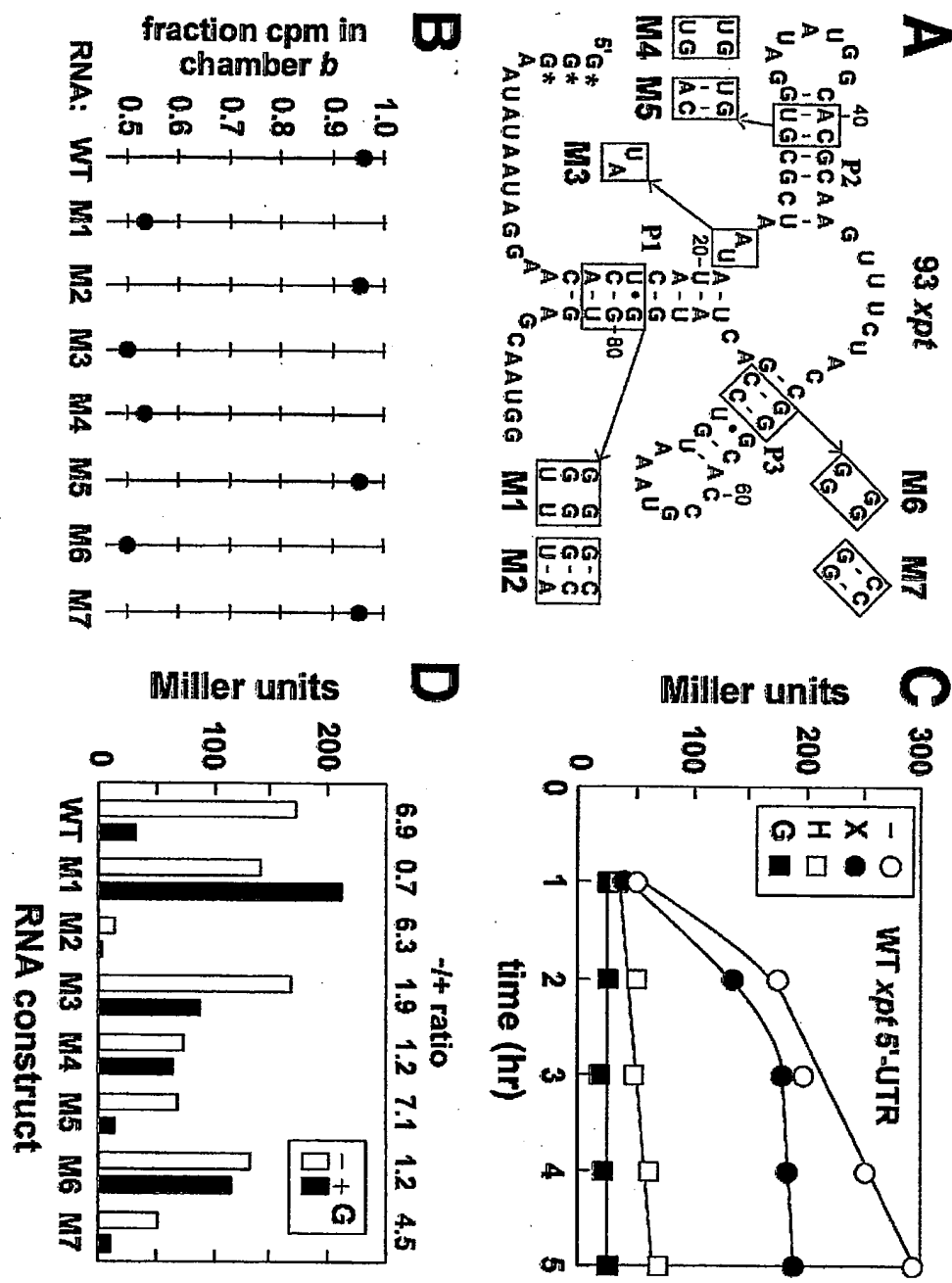
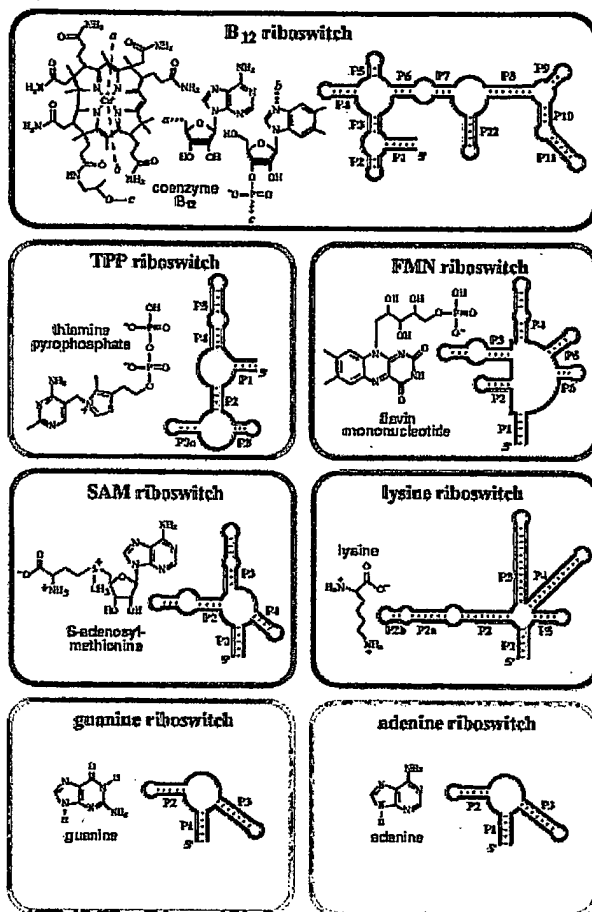


FIG. 28

A



B

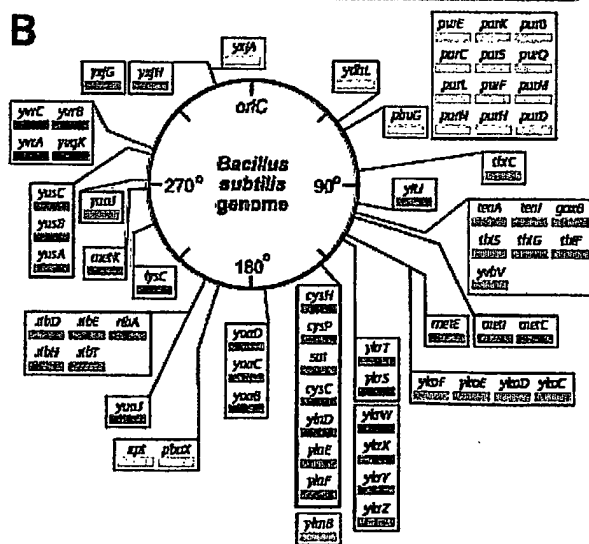


FIG. 29

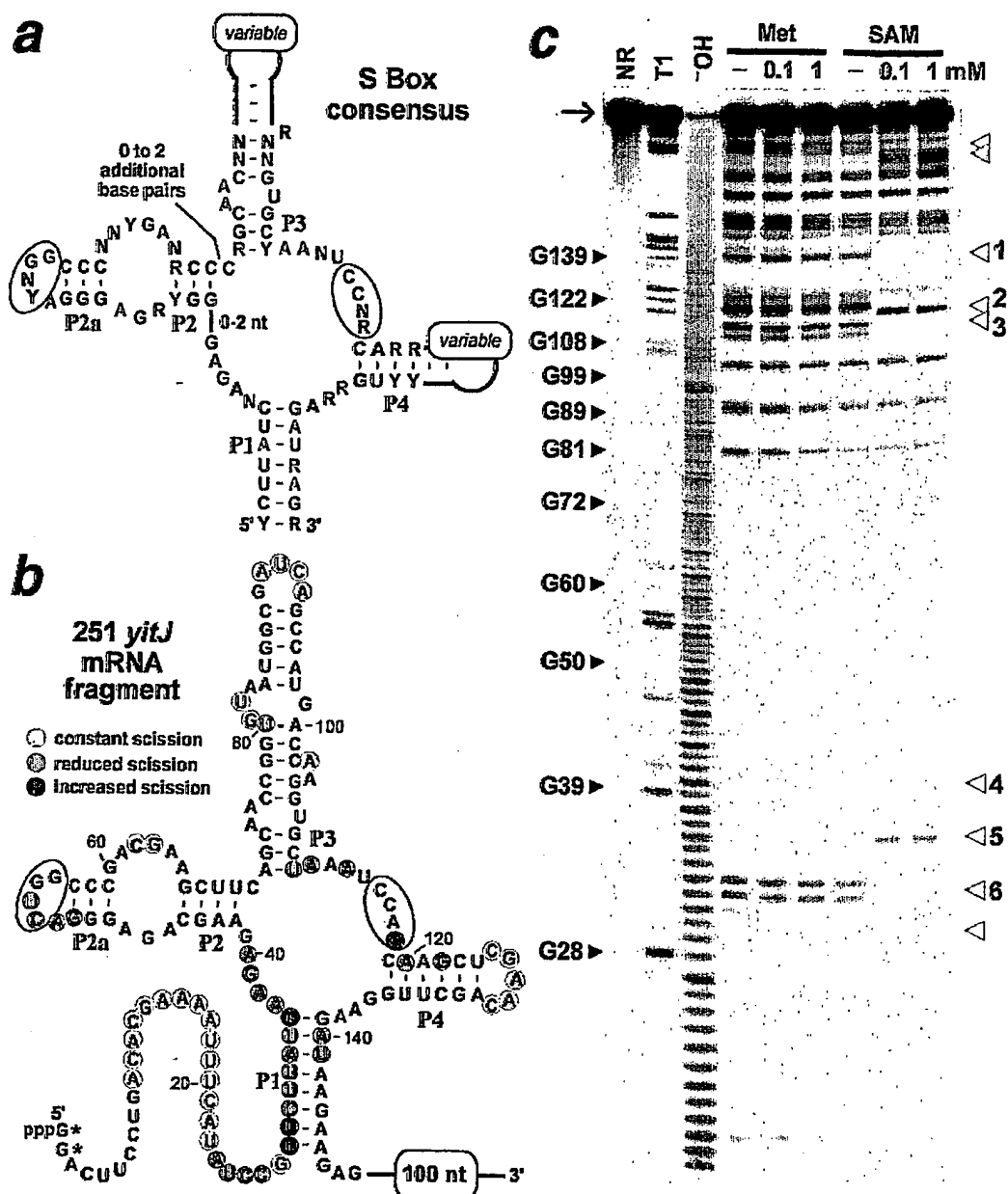


FIG. 30

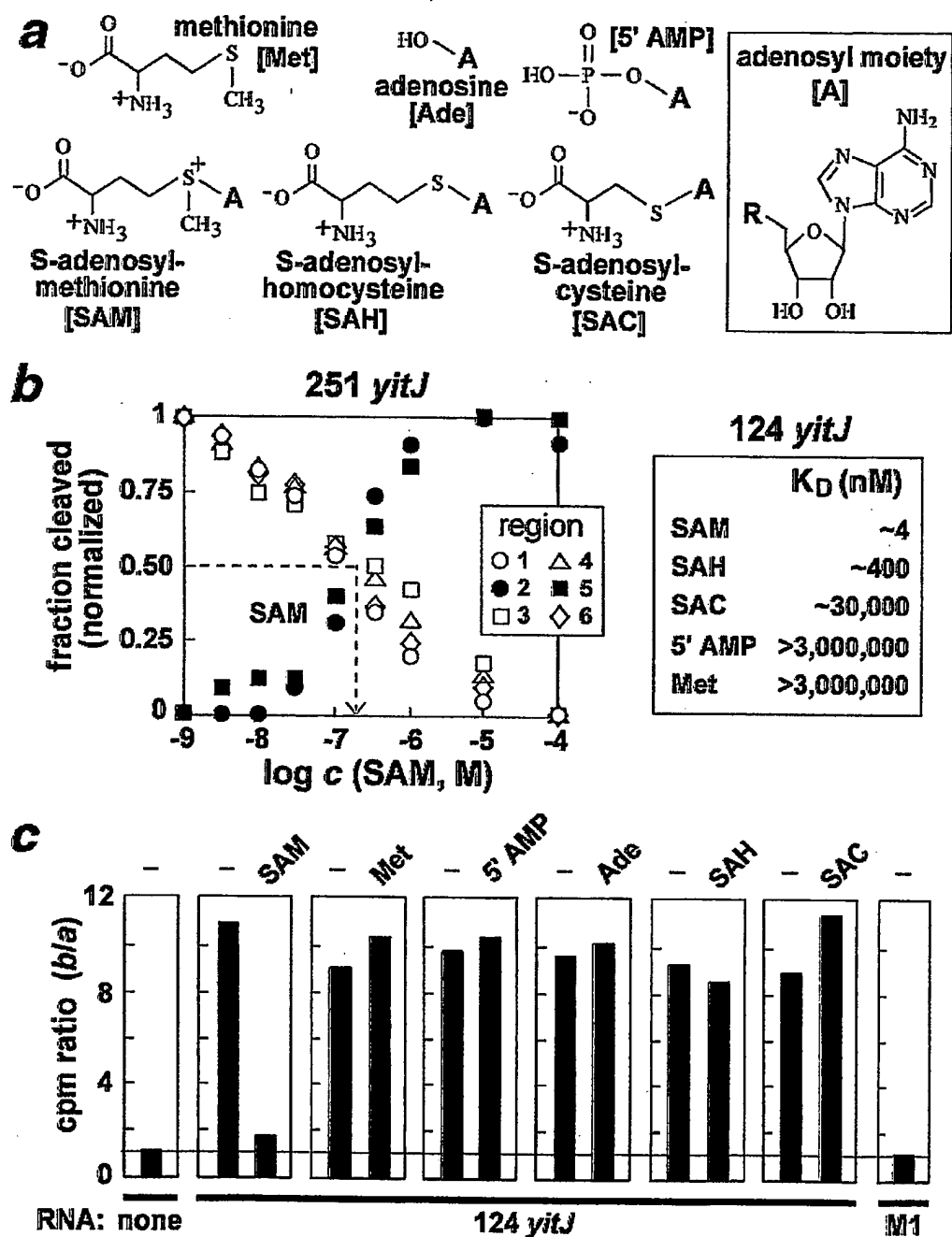


FIG. 31

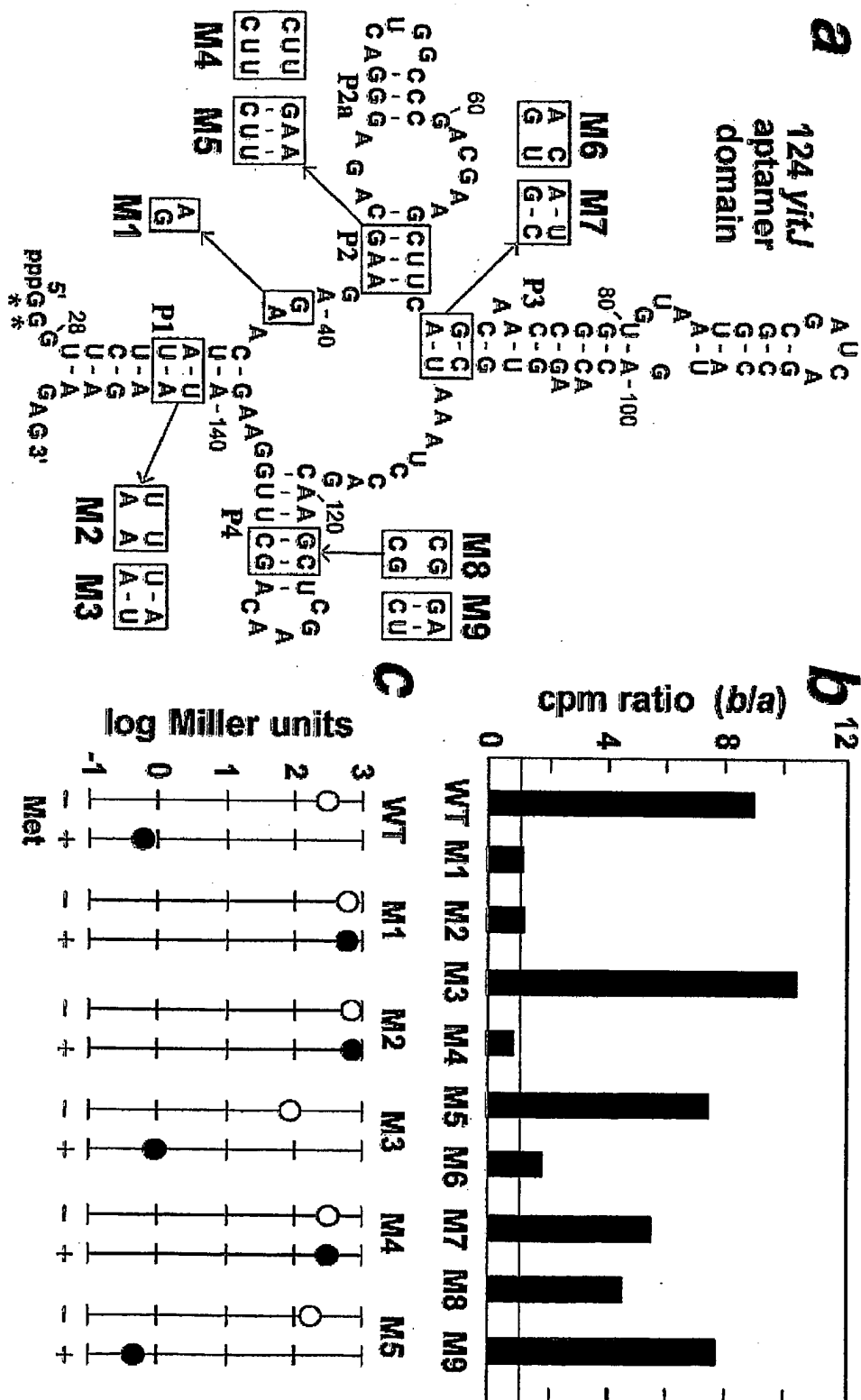


FIG. 32

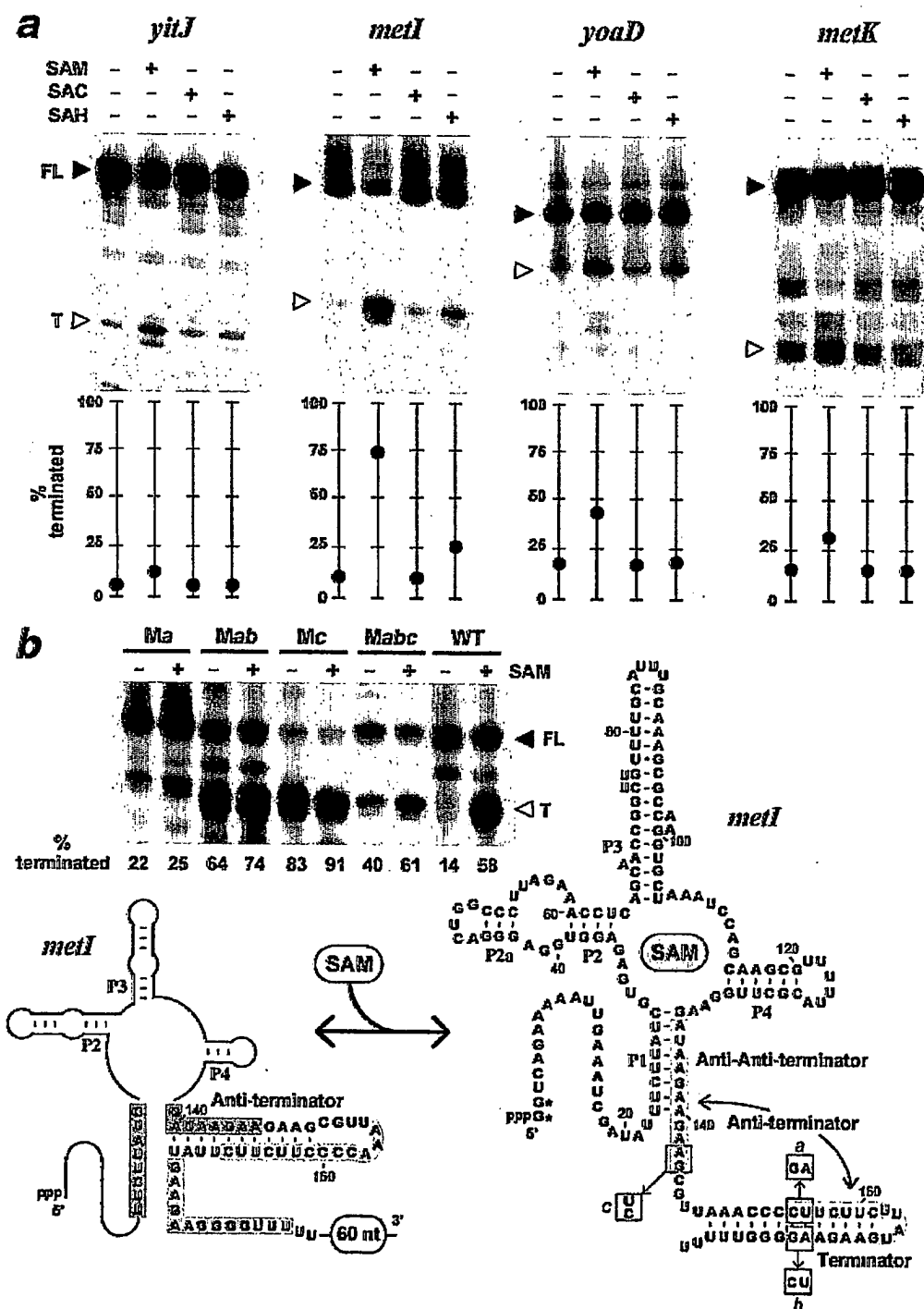


FIG. 33

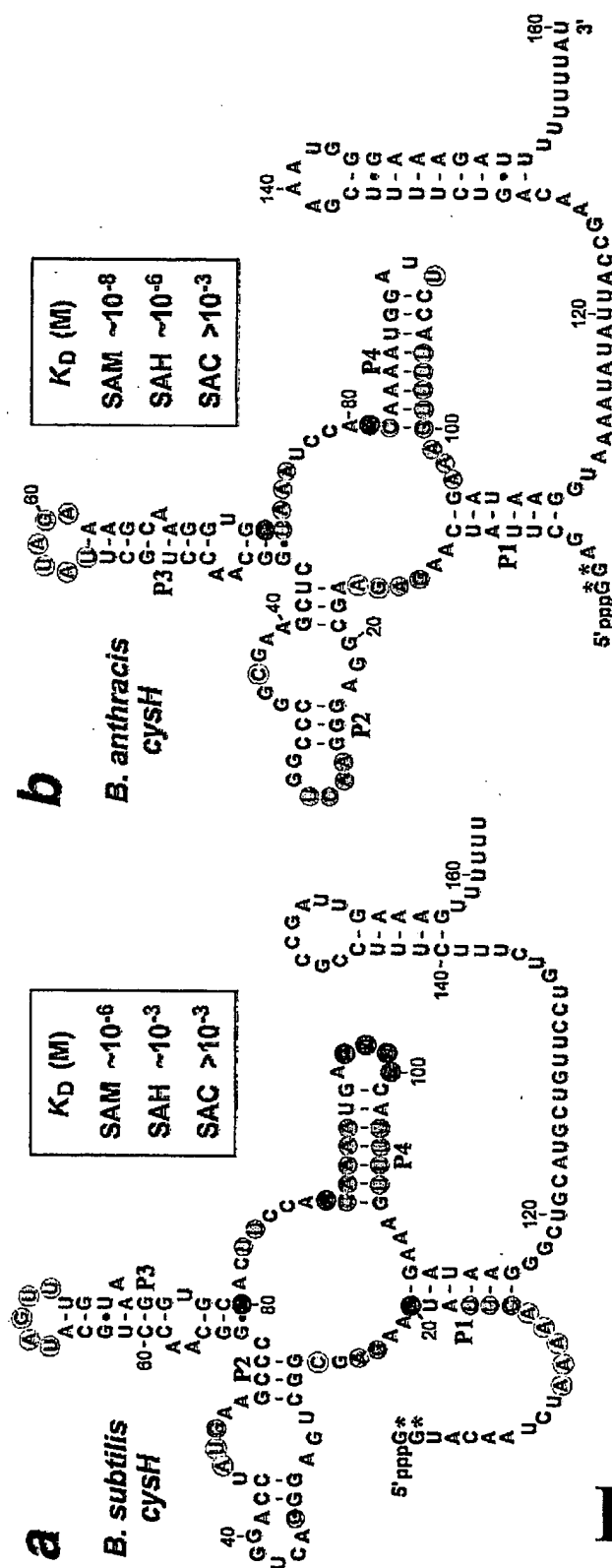


FIG. 34



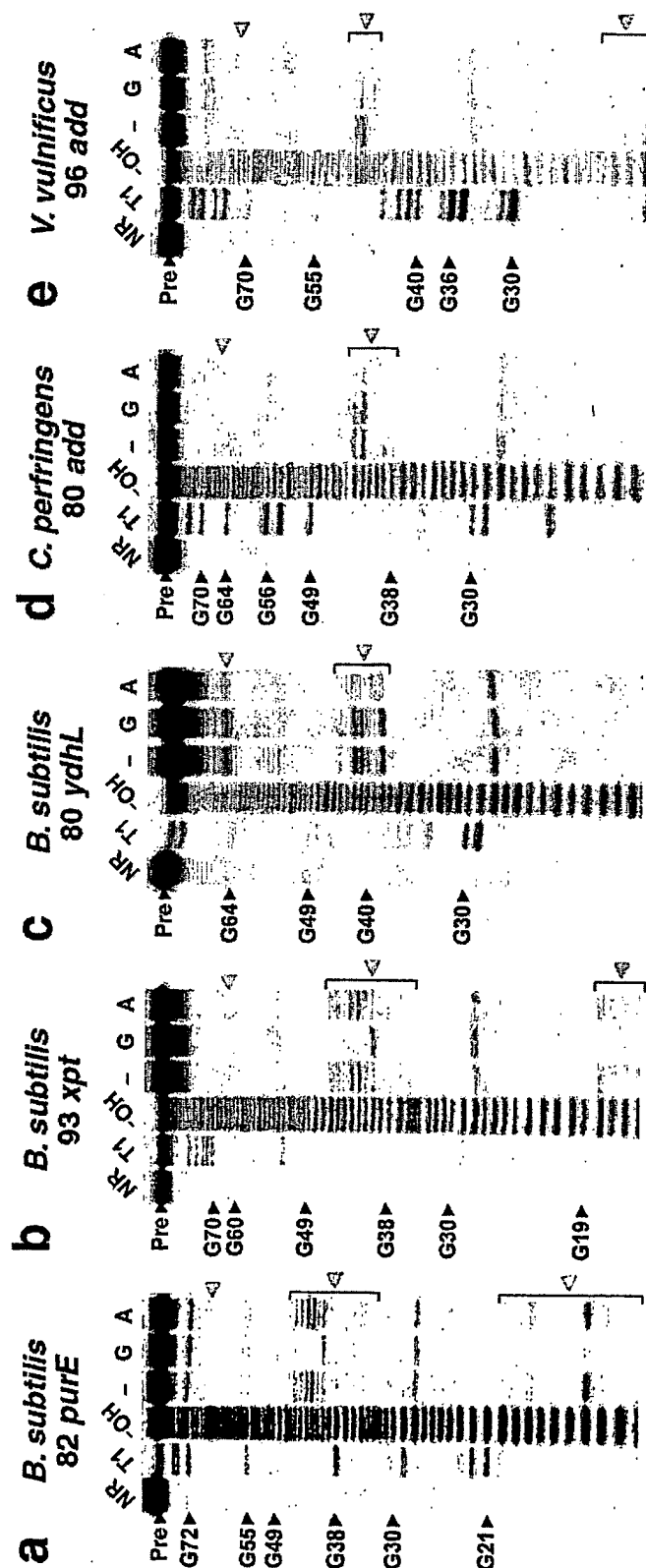


FIG. 36

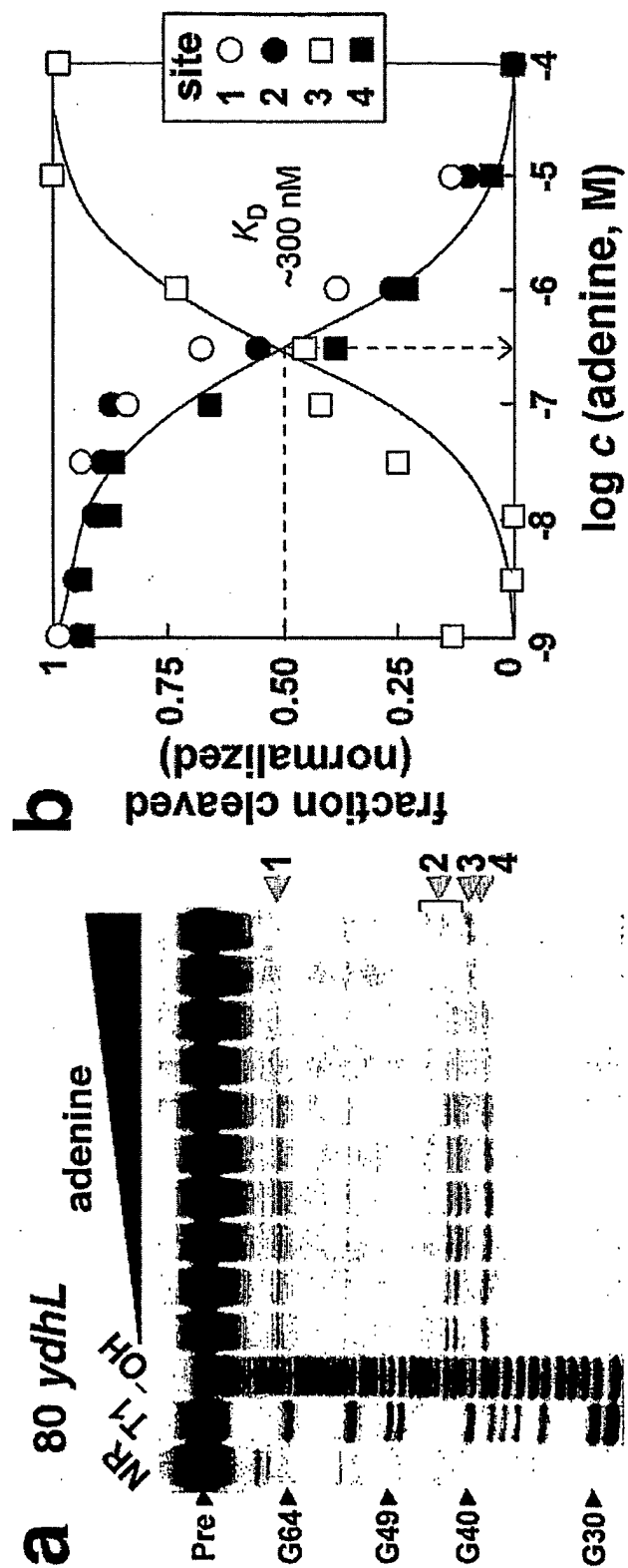


FIG. 37

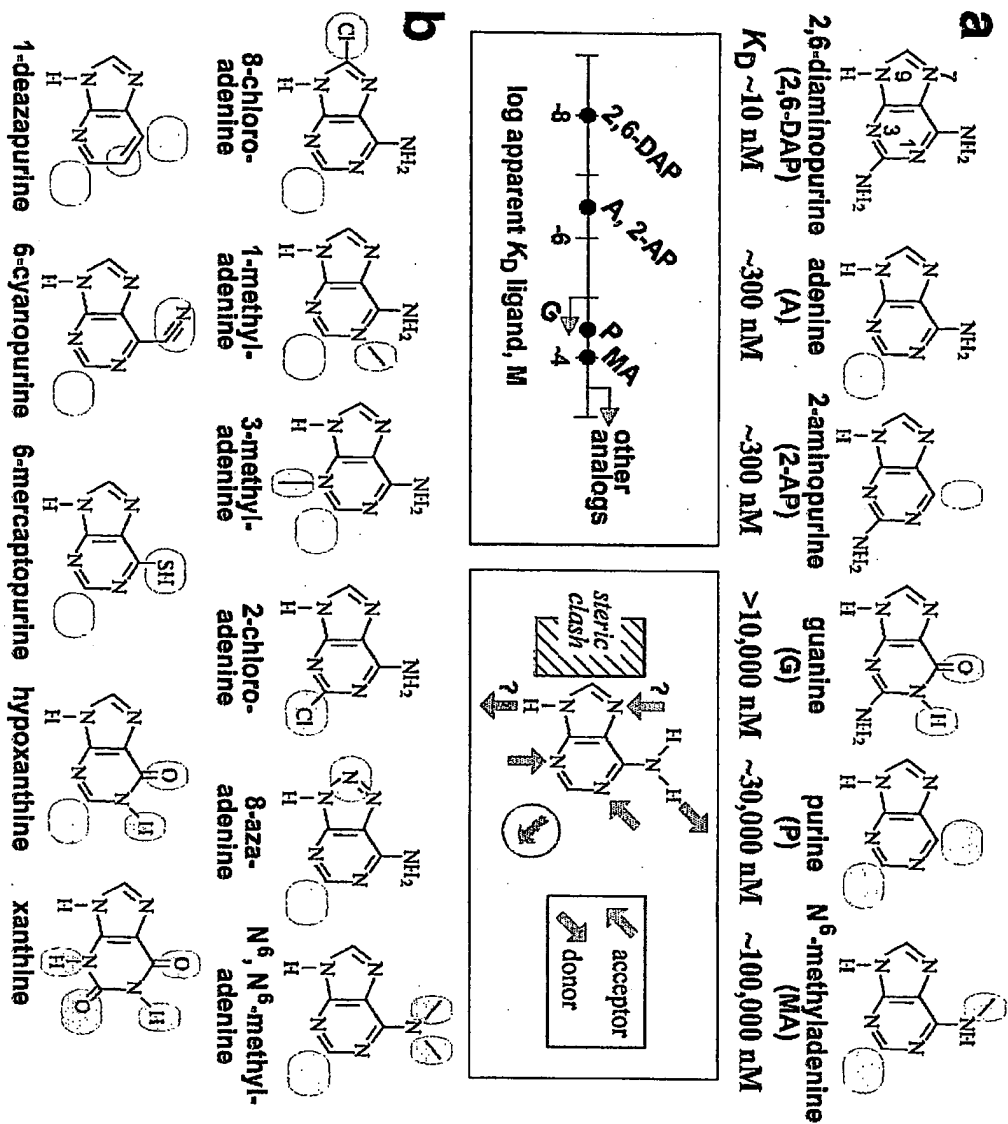


FIG. 38

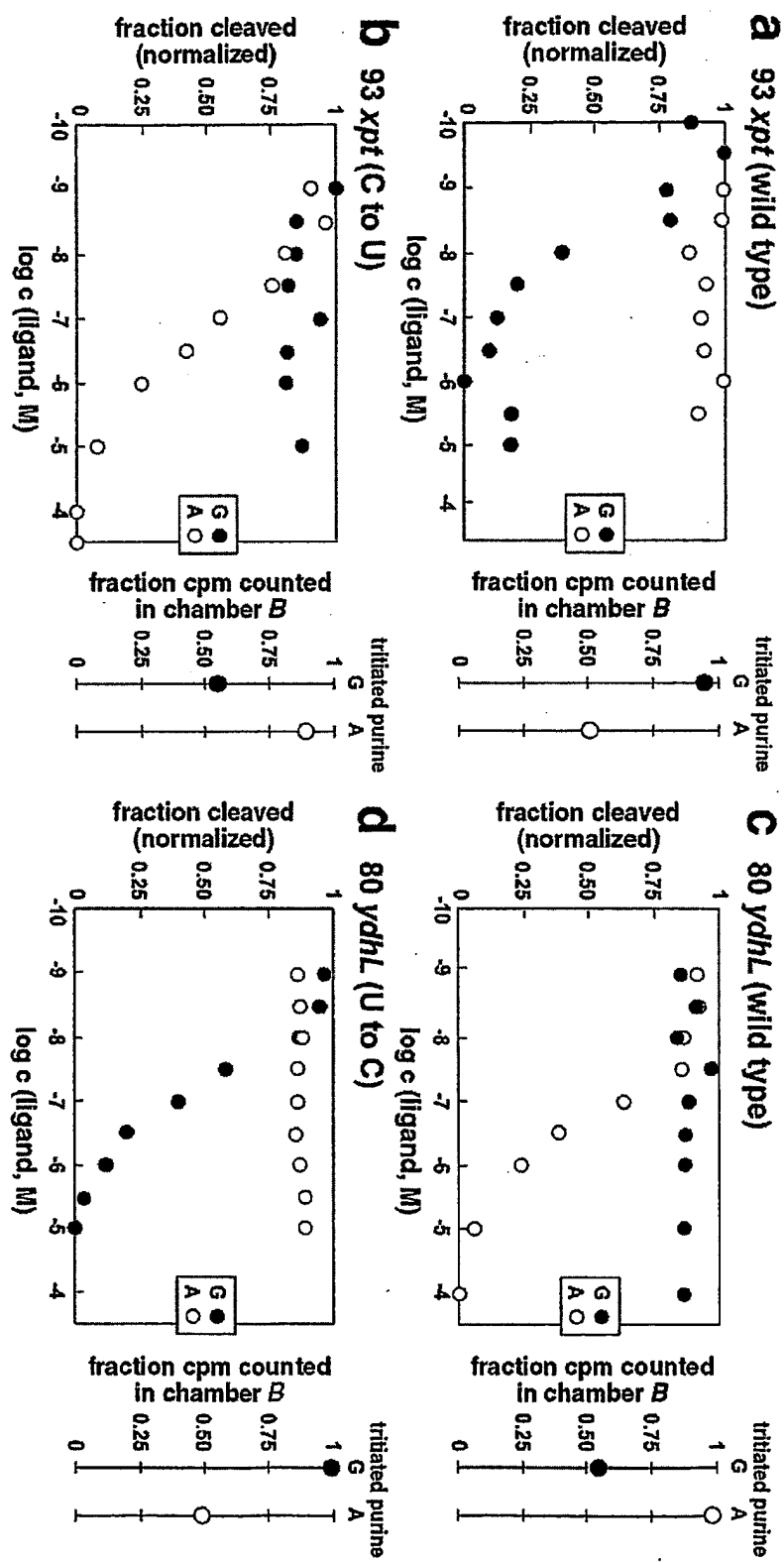


FIG. 39

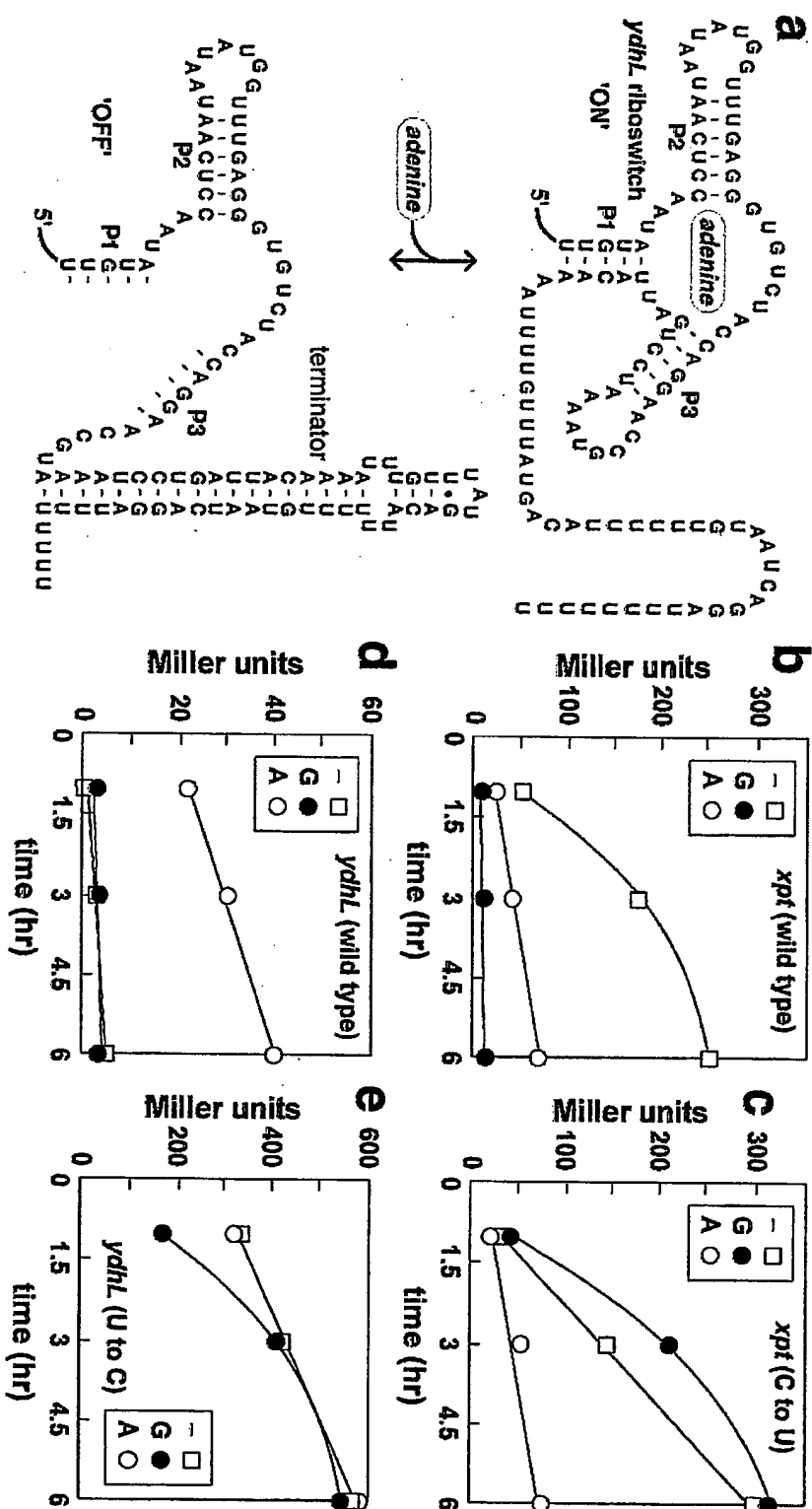


FIG. 40

Figure 41.

A. Alignment of SAM Riboswitches.

Table S1. S Box Sequence Alignment

ID	Position	Genbank Acc.	Organism	Remark	Start Operon	
Bs01	- 1180274	NC_000964.1	Bacillus subtilis		92 metF	(ytlJ)
Bs02	+ 1257777	NC_000964.1	Bacillus subtilis		70 metB-metC	(yjcI)
Bs03	- 1385353	NC_000964.1	Bacillus subtilis		130 metE	(metC)
Bs04	- 1424147	NC_000964.1	Bacillus subtilis	(*)	89 ykrT-GCN3	(ykrT)
Bs05	+ 1426344	NC_000964.1	Bacillus subtilis		60 rbcL-ykrX-araD-ykrZ	(ykrW)
Bs06	+ 1629516	NC_000964.1	Bacillus subtilis		164 cysH-pitA-MET3-cysC	(cysH)
Bs07	- 2024504	NC_000964.1	Bacillus subtilis		86 ldhA-xylB	(yoad)
Bs08	- 3128412	NC_000964.1	Bacillus subtilis		170 metK	(metE)
Bs09	- 3363560	NC_000964.1	Bacillus subtilis		108 abc -2011-nlpA	(yusC)
Bs10	+ 3996569	NC_000964.1	Bacillus subtilis		85 metE	(yxjH)
Bs11	+ 3997959	NC_000964.1	Bacillus subtilis		80 metE	(yxjG)
Bh01	- 910190	NC_002570.1	Bacillus halodurans		141 ????	
Bh02	+ 1348818	NC_002570.1	Bacillus halodurans		99 thrA	
Bh03	+ 1699959	NC_002570.1	Bacillus halodurans		175 metB-metC-metF-metH	

Bh04 - 3427466 NC_002570.1	Bacillus halodurans	157 metK
Bh05 - 3591166 NC_002570.1	Bacillus halodurans	220 abc -2011-nlpA
Oi01 + 727028 NC_004193.1	Oceanobacillus iheyensis	78 meth
Oi02 - 1098097 NC_004193.1	Oceanobacillus iheyensis	162 metE
Oi03 + 1319043 NC_004193.1	Oceanobacillus iheyensis	76 ???
Oi04 - 2134364 NC_004193.1	Oceanobacillus iheyensis	56 abc -???-nlpA
Oi05 - 2365511 NC_004193.1	Oceanobacillus iheyensis	176 metK
Oi06 - 2437305 NC_004193.1	Oceanobacillus iheyensis	129 abc -???-nlpA
Oi07 - 2708643 NC_004193.1	Oceanobacillus iheyensis	177 tran-MET17
Oi10 - 3200636 NC_004193.1	Oceanobacillus iheyensis	81 ???-???
Oi08 - 2856863 NC_004193.1	Oceanobacillus iheyensis	201 ???-???
Oi09 + 3162075 NC_004193.1	Oceanobacillus iheyensis	105 MET17
Oi10 - 3200636 NC_004193.1	Oceanobacillus iheyensis	81 ???-???
Oi11 + 3200766 NC_004193.1	Oceanobacillus iheyensis	94 ???
Oi12 + 3294474 NC_004193.1	Oceanobacillus iheyensis	97 abc-2011-nlpA-abgB
Oi13 - 3466518 NC_004193.1	Oceanobacillus iheyensis	112 gldA-nlpA-abc-2011
Ca01 + 453565 NC_003030.1	Clostridium acetobutylicum	78 metB-metC
Ca02 - 671354 NC_003030.1	Clostridium acetobutylicum	77 meth
Ca03 - 1073886 NC_003030.1	Clostridium acetobutylicum	(smtA-metB-cysK)
Ca04 + 1131539 NC_003030.1	Clostridium acetobutylicum	81 abc-2011-nlpA

(1)

(2)

Ca05 + 1976373 NC_003030.1	<i>Clostridium acetobutylicum</i>	102	metA
Ca06 - 2914839 NC_003030.1	<i>Clostridium acetobutylicum</i>	117	CAC5
Ca07 - 2991405 NC_003030.1	<i>Clostridium acetobutylicum</i>	70	metK
Cp01 - 2500081 NC_003366.1	<i>Clostridium perfringens</i>	391	metK
Cp02 - 2665229 NC_003366.1	<i>Clostridium perfringens</i>	102	nhaC
Lm01 + 137135 NC_003210.1	<i>Listeria monocytogenes</i>	90	oppA-dppB-dppC
Lm02 - 309383 NC_003210.1	<i>Listeria monocytogenes</i>	113	nlpA-abc-2011
Lm03 - 637924 NC_003210.1	<i>Listeria monocytogenes</i>	111	MET17-MET2
Lm04 + 882772 NC_003210.1	<i>Listeria monocytogenes</i>	97	metE
Lm05 - 1716649 NC_003210.1	<i>Listeria monocytogenes</i>	110	metK
Lm06 - 1739595 NC_003210.1	<i>Listeria monocytogenes</i>	109	metE-metB-metC-metH
Lm07 - 2491174 NC_003210.1	<i>Listeria monocytogenes</i>	93	abc-2011-nlpA
Li01 + 172401 NC_003212.1	<i>Listeria innocua</i>	88	oppA-dppB-dppC
Li02 - 327333 NC_003212.1	<i>Listeria innocua</i>	113	nlpA-abc-2011
Li03 - 636911 NC_003212.1	<i>Listeria innocua</i>	111	met17-met2
Li04 + 871751 NC_003212.1	<i>Listeria innocua</i>	97	metE
Li05 - 1772459 NC_003212.1	<i>Listeria innocua</i>	110	metK
Li06 - 1790189 NC_003212.1	<i>Listeria innocua</i>	109	metE-metB-metC-metH
Li07 - 2538251 NC_003212.1	<i>Listeria innocua</i>	92	abc-2011-nlpA
Sa01 + 15958 NC_002745.1	<i>Staphylococcus aureus</i>	41	met2

Sa02 + 875385 NC_002745.1	Staphylococcus aureus	91 abc-2011-nlpA
Sa03 - 1844603 NC_002745.1	Staphylococcus aureus	108 metK
Sa04 - 2381620 NC_002745.1	Staphylococcus aureus	13 nhaC
Sc01 + 4708438 NC_003888.1	Streptomyces coelicolor	26 thrC-moaD

ID	Position	Genbank Acc.	Organism	Remark	Start Operon
Ct01	+ 606192	AE_006470	<i>Chlorobium tepidum</i>		107 CAC5-MET2
Tt01	+ 500245	NC_003869.1	<i>Thermoanaerobacter tengcongensis</i>		55 metK
Tt02	- 1750367	NC_003869.1	<i>Thermoanaerobacter tengcongensis</i>		66 metF-metH-ebsC
Tt03	- 2076680	NC_003869.1	<i>Thermoanaerobacter tengcongensis</i>		78 thrA-CAC5
Ft01	- 987483	NC_003454.1	<i>Fusobacterium nucleatum</i>		84 metK
Ft02	- 1317650	NC_003454.1	<i>Fusobacterium nucleatum</i>	(*)	91 abc-2011-nlpA
Dr01	+ 1363063	NC_001263.1	<i>Deinococcus radiodurans</i> , chr 1		156 abc-2011-nlpA-nlpA
Dr02	+ 980704	NC_001263.1	<i>Deinococcus radiodurans</i> , chr 1		41 metH-????-metF
Xa01	- 3558018	NC_003919.1	<i>Xanthomonas axonopodis</i>		74 MET2-metC-thrA
Xc01	- 3379769	NC_003902.1	<i>Xanthomonas campestris</i>	(*)	73 MET2-metC-thrA
Se01	+ 574	AF_269983.1	<i>Staphylococcus epidermidis</i> genomic clone		
Se02	- 142	AF_270301.1	<i>Staphylococcus epidermidis</i> genomic clone		(*)
Gs01	+ 342843	contig:2947	<i>Geobacter sulfurreducens</i>		
Gs02	+ 2470946	contig:2947	<i>Geobacter sulfurreducens</i>		
Ba01	- 177272	contig:6615	<i>Bacillus anthracis</i>		
Ba02	+ 185586	contig:6615	<i>Bacillus anthracis</i>		
Ba03	- 197185	contig:6615	<i>Bacillus anthracis</i>		
Ba04	+ 320607	contig:6615	<i>Bacillus anthracis</i>		

Ba05 - 371127 contig:6615 Bacillus anthracis
Ba06 + 1362659 contig:6615 Bacillus anthracis
Ba07 + 1375353 contig:6615 Bacillus anthracis
Ba08 + 2459362 contig:6615 Bacillus anthracis
Ba09 - 2953226 contig:6615 Bacillus anthracis
Ba10 - 3091676 contig:6615 Bacillus anthracis
Ba11 - 3890736 contig:6615 Bacillus anthracis
Ba12 + 3892933 contig:6615 Bacillus anthracis
Ba13 + 4074285 contig:6615 Bacillus anthracis
Ba14 - 4074078 contig:6615 Bacillus anthracis
Ba15 - 4553682 contig:6615 Bacillus anthracis
Ba16 - 4739975 contig:6615 Bacillus anthracis (*)
Ba17 - 5140322 contig:6615 Bacillus anthracis (*)
Bc01 + 26115 contig:1617 Bacillus cereus (*)
Bc02 + 748841 contig:1617 Bacillus cereus (*)
Bc03 + 1183078 contig:1617 Bacillus cereus (*)
Bc04 + 1195849 contig:1617 Bacillus cereus (*)
Bc05 + 2200884 contig:1617 Bacillus cereus (*)
Bc06 - 2684484 contig:1617 Bacillus cereus (*)
Bc07 - 2963724 contig:1617 Bacillus cereus (*)

Bc08 - 2773209 contig:1617	Bacillus cereus	(*)
Bc09 + 3500608 contig:1617	Bacillus cereus	(*)
Bc10 - 3687209 contig:1617	Bacillus cereus	(*)
Bc11 + 3687417 contig:1617	Bacillus cereus	(*)
Bc12 - 3498410 contig:1617	Bacillus cereus	(*)
Bc13 - 4205859 contig:1617	Bacillus cereus	(*)
Bc14 - 4397125 contig:1617	Bacillus cereus	(*)
Bc15 - 4784934 contig:1617	Bacillus cereus	(*)
Bc16 - 5114094 contig:1617	Bacillus cereus	(*)
Bc17 - 5094322 contig:1617	Bacillus cereus	(*)
Bc18 + 5101784 contig:1617	Bacillus cereus	(*)

[illegible]

[illegible]

Bc15 GGATAC**TCTCTATG**CCGAG**CT**GGCGAGGGGA**CAGG**CCCGATGAAGCC**CAGCAACCTACGTTGTA**GTGGTAAA
 Bc16 CTGA**TTCTCTATG**AAGA**GA**CGTGGAGGGAC**TGTG**CCCGTGTGAAGCC**CGGCAACCTCAAG**TTAT
 Bc17 TTGCATAG**TCTCTATG**AAGA**GA**AGGTGGAGGGGA**CAGG**CCCGATGAAGCC**TGGCAACAGCGGT**ATA
 Bc01 CGATACAT**TCTCTATG**CAGAG**G**AGGTGGAGGGGA**CTGG**CCGTACGATACCT**CAGCAACCGGT**TTTT
 Bc02 CAACAAT**TCTCTATG**TTGAG**G**AGGTGGAGGGGA**GGGG**CCGTATGAAGCC**TGGCAACCTGCT**ATCAG
 Bc03 AATACAAAC**TCTCTATG**AAGAG**G**AGCGGAGGGGA**CTGG**CCCGCGCAAGCC**GGCAACCTGCT**ATAGA
 Bc04 TGAAC**TCTCTATG**AAGAG**G**AGCGGAGGGGA**CTGG**CCGTACGATGGCT**CGCAACCTGCT**GATTCA
 Bc05 AATTAAT**TCTCTATG**CAGAG**G**AGGTGGAGGGGA**CTGG**CCGTATGAAGCC**CAGCAACCTATA**TATAT
 Bc06 ATGAAT**TCTCTATG**ACGAG**G**AGGTGGAGGGGA**CTGG**CCGTATGAAGCC**CGCAACCTATGG**TTA
 Bc07 ATTAG**TCTCTATG**AAGAG**G**AGGTGGAGGGGA**CTGG**CCCGATGAAGCC**CAGCAACCTGCT**ATAAA
 Bc08 GAATAT**TCTCTATG**CAGAG**G**AGGTGGAGGGGA**CTGG**CCCGATGAAGCC**CAGCAACCTGCT**GAT
 Bc09 TAAATAC**TCTCTATG**AAGAG**CA**CGTGGAGGGGA**CTGG**CCCGATGAAGCC**CAGCAACCTGCT**AATT
 Bc10 AGACAAAC**TCTCTATG**CAGAG**G**AGGTGGAGGGGA**CTGG**CCGTATGAAGCC**CAGCAACCTGCT**GAAT
 Bc11 ACGAAT**TCTCTATG**TAGAG**G**AGGTGGAGGGGA**CTGG**CCGTATGAAGCC**CAGCAACCTGCT**ATT
 Bc12 TATACAA**TCTCTATG**AAGAG**CA**CGTGGAGGGGA**CTGG**CCCGATGAAGCC**CAGCAACCTGCT**GTAAAT**CGAATGTGAATGCGCGCTTTATTTCGCG**
 Bc13 GGATAC**TCTCTATG**CCGAG**CT**GGCGGAGGGGA**CAGG**CCCGATGAAGCC**CAGCAACCTGCT**ATTGGTAAAC
 Bc14 CTGA**TTCTCTATG**AAGA**GA**CGTGGAGGGGA**CTGG**CCCGTGTGAAGCC**CGGCAACCTGCT**TTAT
 Bc15 TTGCATAG**TCTCTATG**AAGA**GA**AGGTGGAGGGGA**CAGG**CCCGATGAAGCC**TGGCAACAGCGGT**ATA
 Bc16 TTTA**TCTCTATG**ATGAGAG**G**AGGTGGAGGGGA**CTGG**CCGTATGAAGCC**CGGCAACAGCTTGA**TTTT
 Bc17 ACACATAC**TCTCTATG**AAGAG**GT**CGCGGAGGGGA**CTGG**CCCGATGAAGCC**CGGCAACCTGCT**ACG
 Bc18 AGCAAT**TCTCTATG**CAGAG**G**AGGTGGAGGGGA**CTGG**CCGTATGAAGCC**CAGCAACCTGCT**GTAATA

Cons t. GTCCYaa.t.Cc.t yylgmGATRgr.tr
 B301 GCGATGACGAAAGGTGCTAAAT-CCAG-CAAGCT-CGAACA-GCTGGAGATAAGAGAGACAAAATCACTGACAAA-
 B302 GGAAGGCGGAAAGTGTGCTAAAT-CCAG-CAAGCGT-TTTT-ATGCTGGAAAGATAAGCAAGAGCGTTAA-
 B303 TGATGTCTATAA-GGCACGGTGTCTAAAT-CCAT-CAGAT-TCT-GTCTGAGATGAGAGAGCAGTGTTTTACGTAGAAAA-
 B304 GCTGCTCTTAA-GGCACGGTGTCTAAAT-CCAT-T-CAGAT-CTG-ATCTGAGATAGAGAGGCGGACATAGATGTTAA-
 B305 AAGCAGCGTGTCTAAAT-CTTG-CAGCT-AGC-AGGTGAGATAGAATTCGGACGAGAAAAAGAAA-
 B306 TCTTAAGGTGCT-ACTTCCAG-CAAAATG-AAATTC-CAATTTGAAAGATAAGGCTGCATGCTGTTCCTGT-
 B307 TGTGCGCAAT-CCAG-CAAGC-GCTA-GCTTAAAGATAGAAAAGCAAGGTTTATACCGCGCTGCGCTG
 B308 GAGACAGCAAGGTGCTAA-CCTG-TTCCAAAGG-TTGATGATT-CTTTAGCGATAAGAGTGAAGGCACAAAGACCAA-
 B309 GTTGAATATGGTGGCAAT-CAGA-CGAAGC-GTTCA-TCCTTGAAGATGAGAGAAAGCATTTTATATA-
 B310 AACAGAAATGGTGTAAAT-CCAT-AAGAC-ATTGC-GTTCTGCAGATGAGCGGAGATTGTATCGTTCAA-
 B311 AGCAATGGTGTAAAT-CCAT-TAGAGCA-TGA-TTGGCTTGAAGATAAGTTGAGATTGTACGCAA-
 B312 AAGAGAGGTGCCAAT-CCAG-CAGAACA-TGAC-TGATGTGAGATAGCGAAGCGAAGGCGAAAA-
 B313 AGCAACGAAAGGTGTAAAT-CCAG-CAAAATC-ATT-CAATGTGAGATAGAGGAAAGCGAGTGAGATCCAA-
 B314 TGTGGTCAAGGTGTAAAT-CCAG-CAAGCA-TTATT-TGCTGAGATAGAGGAAAGCGAGTGAGATCCAA-
 B315 AGGTGAAAAGGTGTAAAT-CCAG-CAAGGC-GTT-GCTTGAAGATAGAGCGGAAAGGTAATGTTAATAA-
 B316 GTTGGAAAAGGTGTAAAT-CCAG-CAAGC-GAT-GCTTGAAGATGAGAGAGGAAAGCGTAAACAT-
 B317 AGGTATGGTGTAAAT-CCAG-TAGCT-TACA-CAGCGTTAAGATAGAGAGAGCTATGTATTTAA-
 B318 TAAGTATGTGTCAAT-CCAG-TAGCG-TAAT-TGCTAGAGATGAGAGAGTATATAGTACGGTTCCTGTGA-
 B319 TGTGTGTAAAT-CCAG-CAAGC-AATA-GCTTGAAGATAGAGTGTATGCTAAATATCCAAATCTCT
 B320 GGAGTGTGTCTCTTCCAG-CAAGAA-TTTT-TTCTGAAGATAGGTAATGATATGTAATAA-
 B321 GTGTGAAAAGGTGTAA-TCTG-ATCGAGCA-TAATAGT-TGCTTGAACATAAGAGCGAAAGGCCATAATCTTT-
 B322 ATGAAAGGTGTCAAT-CCAG-CAAGCA-TAAT-TGCTTGAAGATAGAGTGAAGTAAATTTATAGCTTTAAAA-
 B323 GAAATAGGTGTCAAT-CCAT-CAAGTAT-TCT-ATGCTGTGTAGATAGAGGAAGTGTGCGGACAGAG-
 B324 GTAAAGGTGTCAAT-CCAG-CAAGC-GTAG-GCTTGAATGATGAGAAAAGTGTTTATACCTTTTAAATAAAA-
 B325 GCTATGAAAGGTGTCAAT-CCAT-CAAGC-TTCATC-GCTTGAAGATAGAGAGAGTGTGCTTTTAAACAGACAAA-
 B326 GAAATGTGTCAAT-CCAT-CAAGTAT-TCT-ATGCTGTGTAGATAAGGAAGTGTGCGGACAGAG-
 B327 TAGAAAAGGTGTACCT-CCAG-CAAGAT-GTAT-GCTTGAAGATAGAGTCCAGATTAAAAAAA-

O112 TAAAGAACTGTGCGAATTCCG CAAATGGAAGATGAGAAAGCATGGCTTCTACATATATACATATG
O113 GTGATGAATAGGTGCTAAATCCG CAAATATAC GGACA GTATTTGAGAAATAGAGAGGTGATGAAATGACATTACGTAGTGTAA
Ca01 TTGGAGATGTCTAATTCCG CAGC GTTAT CAGCAGATGAGAAATGTTTTTAA
Ca02 AATATAATGTGGTCTAAATCCG CAGC AAGC GTGATGAGAAATATCGCAATGTAAA
Ca03 ACTATATGTGGTCTAATTCCG CAGCA TATT TCTGAAAGATGAGGAGCGACTATTTAAACATTTTATTTTGT
Ca04 TATACAGAGGTGCTAATTCCG CAGC GCTA TCTGAGAGATGAGAAATATAATCGAGCTTTA
Ca05 GTACGTGTCTAATTCCG CAAAG ITATTT GTTTGAAAGATAGAAACAGCTTATTAATTAATGAGTATGTT
Ca06 AGATGATGTGTCTAATTCCG CAAAG ITAA TTGTGAGAGATAAGAGGATTATAAAATTTTAGAAAGCTAAAA
Ca07 AGTACATAATGTGGTCTAATTCCG CAGAA ITA TTGTGCAAGATAAGAGAGAGAAATGTAA
Ca01 TCACTACGTGTCTAATTCCG TAAAGA AAT TGTGCAAGATGAGAGAGATAAATTTAGTGTATAACTAAAA
Ca02 ATTTCTATGCAAGATTTATGCGGTCTAATCCG AGAA ACTGAGACATAAGAGAGAGCTCTGTAAAGATAATAA
Lm01 GTTCTATGCTAATTCCG CAGAA GTAATA TCTGGCAGATAGTAGTACCTTTCAATGAGG
Lm02 ATAAAGTGAAGTGTCTAATTCCG CAAATGG TGTATT TCGTTTGTAGATAGAGAGCTGGATATGTTGACTTTCC
Lm03 TTTCAGGTGTCTAATTCCG CAG TATATT CTGAAAATATAGTCGGAAATCCAAAGTTTAGGAAACTCTAT
Lm04 TCACGTGTCTAATTCCG CAG TAAACA CTGACAGATAGGCACGCCAATCAGGTAAATTTACT
Lm05 GTGAAAGTGTCTAATTCCG TTTG TATT TTTCTAACGATGAGAGCAAGGTATAATTAT
Lm06 TTGAAAGTGTCTAATTCCG CAAAG TGA TGTGAGAGAGATAAGAGAGAGACTTAAAGTTTCAGTGTATTGT
Lm07 TTATGTTTAAAGTGTCTAAGTCTAG CAGAACAA CTAA TTGTTCTGAAAGATGAGAGGAAGTTAGTCCATTGAAAAAATGCT
Li01 GTTCTATGCTAATTCCG T CAGAA GTAATA TCTGGCAGATAGTAGTACCTTTTAATGAGG
Li02 GTAAAGTGAAGTGTCTAATTCCG CAAATGG TGTATT TCGTTTGTAGATAGAGAGCTGGATATGTTGACTTTCCA
Li03 TTACGTGTCTAATTCCG CAG TATATT CTGAAAATATAGTCGGAAATCCAAAGTTTAGGAAACTCTAT
Li04 TCACTGTGTCTAATTCCG CAGT AIC ACTGACAGATAGGCACGCCAAGGTAAATCACT
Li05 GTGAAAGTGTCTAATTCCG TTTG TATT TTTCTAACGATGAGAGCAAGGTATAATTATA
Li06 TTGAAAGTGTCTAATTCCG CAAAG TGA TGTGAGAGAGATAAGAGAGAGACTTAAAAAGTTTCACTGTATTGT
Li07 TTATGTTTAAAGTGTCTAAGTCTAG CAGAACAA CGAT TTGTTCTGAAAGATGAGAGGAAGTTAGCCCATTTGAAAAAATGCT
Sa01 AGCAAGGTGTCTAATAA CCA TTA TCTGAAATGATAGTATAAGA
Sa02 TAAATGTGTCTAATTCCG TAAAGI ITTA ACTTTGAGATGAGAGAAACAATACTACTAT
Sa03 AAAGAAAGTGTCTAATAA CCGI TTGAGAG TTTG TAAAG CATAAGAGCGAATGAGACGTATTA
Sa04 ATGTGCTAATTCCG TAAAG TAA TTGTTTGAACATAAGCAGGTAAAGCACATGAAA
Se01 CCGCTACGAAAGTGTCTAATCCG CAGAG AGATG CTTGCTGAGAGAGATGAGAGGAAGGCGCTCCGCTCCATGGCTGTGC

ConsT. GTGCYaa.t-Cc.r.....catt.....yyigraGATRagrr.rr.....

C01 ATGAGCGGTGCAAAAT CCAAT CCGGGA GGAAG TCCGGGAAAGATGATGTATGCATTCTCTGCTGATTTCACTCACTCA
 T01 GGAATAGGTGCCAATT GTCT CAGAAGA TTTT TCTCTGAAAGATGAGGGTATG
 T02 GGCATGTGTCCTAATT CCTG CAGCG GTT CCGTGAAGATGAGAGATCTTGTA
 T03 GTGCTGTGTGCCAATT CCTG CAGGTGGG GTTA CCGAGCTGAGATGAGAGGAGAGCGCGAGTAATTGTGA
 F01 TGTGTGTGTGCTAATT CCTG ATAGATGCAAAAGATTATAATACATCT
 F02 GTGTGTGTGCTAATT CCAG AGAGATCGAGAGGAAATTTGAAACAAGAACTAA
 D01 GGCAGGGGTGCT TT CCAG A GCGCGCGACACAGGCGCGACATGCGCGCGCGCGGGAACGATAAAGGAAGCGGGGTCTCTTCGGGGTTCCAA
 D02 GGTCACTGGTCT TT CAGGAAA GCGCGCTTTAGTCTGCGGACGATGCGCGGAG CCGCG CATTG CCGCC CAGGAGGTGCATTTCACACCCATTGAGC
 Xa01 GCGTCGGGTTCGTGCGAATT CCTG CCGGGAG CTCCGC CTCGCGGAAA GATGGTTCGAATCTGCGCTTGCGCACGTCGAAACGC
 Xc01 GCGCGCGGTTCGTGCGAATT CCTG CCGGGAG CTCCGC CTCGCGGAAA GATGGTTCGAATCTGCGCTCTGCACGTCGAAACGC
 Se01 AGCACTGTGCTAATA CCA GAG CAA CTCTAATATATAGTACGATAA
 Se02 AAGAAGGTGCAAA CCGT TTCTGAGC GCTGTGAACGATATAGAGCGAATGAGCTTTTAAGAGCCTTCTCTC
 Gs01 CCGACCGCACTGTGCTAATT CCTG CCG CAG GAA GCGAGCGATGAGAGGAGCTCTGTGACCAACCGACGCGTACA
 Gs02 GCGGAAAGGTGCAATT CCTG CCGAGAC GACA GCGTTCGGGAGATAGGAAGCGGTGACACCTCACGGTGAATCGA
 Ba01 TATAAGCTAAGGTGCTAATT CCTG CAAATAG GATTTCGGGAGATATAGAGAGGATCCTATTTTGTCTATTCTG
 Ba02 GGAACACCTGTGCTAATT CCAG CAAG CAAGT CTGAAA GATTAAGTATGCGCCCTTGTTTATTAA
 Ba03 TGAATAGTGTGCACTT CCTG CAAAGCT TTAT AGCTGAAA GATGATGAGAGGAGCTCTGTTTATATACGGGTGCA
 Ba04 AATACCGTGTGCTAATT CCTG CAAAGC ATATAA GCGTTCGAA GATGAGAG GATGTGACCGAGTACATATA
 Ba05 AAGCTGAATGGTGTGCTAATT CCTG CAAATAGC ATT CCAATTGAAA GATATAAAGCTGACTATTGTGTACAAA
 Ba06 AAGCACTGTGCTAATT CCTG CAAAGG AAT CCAATTGAAA GATATAAAGCTGACTATTGTGTACAAA
 Ba07 GAGTGTGCGCAATT CCTG CAAAGC ATGT GGTGAAAAGATGAGAGAGCGCTTCTTATAGATGTATA
 Ba08 GCAATGGAAGGTGCTAATT CCTG CAAAGAGAGC GTT GGTGTTTGGGAGATGAGAGGATCTTGAACGTGAAAGAAA
 Ba09 TGAATAGTGTGCTAATT CCTG CAAAG GTAA CTTGAAAAGATAAGAAAGAGCTCATTTTGACTATATATACAG
 Ba10 GGAAGTGTGCTAATT CCTG CAAAGCA AATT TGTGTTGGGAGATATAGCAAGATATATACGTAA
 Ba11 AAA GCGACGGTGTGCTAATT CCTG CAAAGAGT AAA ACTTGTGCGGAGATATAGAGGGGAGAGATAAACTTCAA
 Ba12 TGTAGACAGGTGCTAATT CCTG CAGG ATTAC GCTGACAGATTAAGGAGCTGTGTTGTAAAAAAA
 Ba13 GTTAAAGGTGCTAATT CCTG CAAAT GCGAAA AATTTGACAGATGAGAGGAGAGCTCTATTTCACACCGAAA
 Ba14 GTTGAAGCGGTGCTAATA CCTG CAAAG GAAT GTTGTGCTAATAAGAGGAGGAGAACTTATGTT

Ba15 TACAGGTGAATAGGTGCTAAAA-CCIG TG-CGAGGCT ACA- BCTCTCGAACGATAGAGCGGAAGGCAAAAGACGATATGCAAAGTA
Ba16 GTTGAATATGGTCCCAATT-CCIG CAAAGC AAATG CTTTGAAGAGATGAGAGAGGGGATAATGTTGTTATATACGCAT
Ba17 AGGCAATGTGTGCAAAAT-CCIG CAGG TAATAAT CTTGAGAGATAAGAAAGAGCCTTTAGAGCGTGTTCCTCAA-
Ba01 AATACCGTCTAACT-CCAG CAAAGCT ATGA AAGCTTGGAAAGATGAGAA GATGTGAACGAGTACATATA-
Ba02 ACGAAAGGTGCAAAAT-CCIG CAGTG AAGAAA CAGCTGAAAGATAAGAGCGGTCAATTAGTCAAGAA-
Ba03 AAGCAAGTGTCTAAAT-CCAG CAAATGG AAT CCAATTTGAAAGATAAGGTAAATAATATTACCGAACAG-
Ba04 GATGCTGTGCAAAAT-CCAG CAAAG ATAT CCTTAAAGATGAGAAAGAGCGTTCCTTCTATAGATGTATA-
Ba05 TATAGGAAGGTGCTAAAT-CCG CAGAGACAG GAT GTTCTTTGGAAAGATAAGAGGATTCCTTGAACGTGAAAGAAA-
Ba06 TGAATACTGTGCAAAAT-CCAG CAAAG GTAA CTTGAAAGATAAGAAAGAAAGCTCATTTTGAACGTGATATGACAG
Ba07 AGTACGTGTCTAACT-CCAG CAAAGGT ATGA CCGTTGGAAGATGAGGGGAATGGATTAAACATTCAA-
Ba08 CCAAGGTCTAAAT-CCAG CAGAGCA TATT TGTCTGGAGATAAGACGAAGATATATACGTAA-
Ba09 GTAGACAGGTGCTAAAT-CTCG CAGC ATTAC CTTGACAGATAGGAGCTGTTGTAAAAAA-
Ba10 GTTGAAGACGTGTCTAAATA-CCIG CAAAGC GAAT GTTTGCAATATAAGAGGAGGATCGATTATGT-
Ba11 GTTAATAAGGTGCTAAAT-CCAG CAAAT GTGAAA GATTGACAGATGAGAGGAAAGACTCTATTCAAACCGAAA-
Ba12 AAAA GGCACGGTCTAAAT-CCAG CAGAAAGT AAA ACCTTCTGCAAGATAAGAGGGGAGAGATAAACTTCAA-
Ba13 CCAAGTGAATAGGTGCTAAAA-CCIG TG-CGAGGCT ACA- BCTCTCGAACGATAAGAGCGGAAGGCAAAAGACGATATGCAAAGTA
Ba14 GTTGAATAGGTGCTAAAT-CCIG CAAAGC AAAT CCGTTGAGAGATGAGAGAGAGGGATAATGTTGTTATATACGCAC
Ba15 ACGGAATGTGTGCAAAAT-CCIG CAGG TAATAAC CTTGAGAGATAAGAAAGAGCCTTTAGAGCGTGTTCCTCAA-
Ba16 TGAATAGTGTGTGCACTT-CCIG CAAAGC TTAT AAGTTGAAAGATAGAAAGAGCGACTTCGTTTATATACGGTGA
Ba17 TATAAGCTAAAGGTGCTAAAT-CCIG CAAAGCGA GTTC TCGTTTGGAAAGATAAGAGAGGAATCTATTGTCTATTG-
Ba18 TGAAGACGGTGTCTAAAT-CCAG CAAAG CAAGT CTTGAAAGATAAGAGATGGCCCTTTGTTTATAA-

Cons.mm. tttt

BS01CTT.....AAGAGACTTTTATTTCTCTTTTCCCTGCTGATGTGAATAAAGGAGGCAGACA

BS02TAT.....GAAGAAGGGTTTTTATTTTGAAAGGAAAGGTGTCAGCTATATGTCACAGCAGTTTGAAA

BS03TCAT.....GGGAAAGAGGGTTTTTGTGTGAGAAAACCTCTTAGCAGCCTGTATCCCGGGTGAAAGAGA

BS04TCT.....GAGAAAGGAGGGTTTTTACGGCCACATATTAATTAATTACATAATTGGAGGTTATGATGATG

BS05GATT.....GCAAGTTTCAAGAGGTTTTTTTGAATGAGATGAAAAGGAGCTGTGTCATGAGTGAOTTTATTA

BS06TCC.....GCAATTGAAGTTTTTTTATTTAAGAGGTAAAGGCTATCTGTATATACGACGCCGCGAAT

BS07 TAACAGAGCGCCTATATGAATGCTTGG.....ATCTTC.....GGAAGAGGATTTTTTTTTATGAAAAAATACGATGAAAAGGATGTTTTCAGCATGACCGTTT

BS08TCCGAT.....GGAAGAGGTTTTTTTATTTTATTTTCATAAATATGCCAAATTAACATTTCTCTAAATATAACTGTACA

BS09TCAAGTGT.....GCAGAAAGGGTTTTTTCGTGCAAAAAAACCAGGAAGATTTCTTAGAAATAGTGTAAAGCAG

BS10ACACA.....AAGGAAGAGGTTTTTACATGCTTAATATTTTCAGAAAAGAGCGGAATAAACATGCTCAACAA

BS11ATCCA.....AAGGAAGAGGTTTTTTTATTTGAATGGAAAGAGGAAATGGCAACATGTCACAACAACA

BH01TTATC.....GAGAGAGGTTTTTCTGTGTTTAAACACCTTATCTGTGCGGAAAGATTACTTCTTATGTGA

BH02TTATC.....GGAAGGTTTTTTTGTGTAGAGCCCAAGTTTTTATACAAATCAGAAAAGATCGAACTTTTCAGATAGTTTAAAGAAAAA

BH03AATCTTACATGAC.....GGAAGGTTAGGTTTTTTTACAAATCAGAAAAGATCGAACTTTTCAGATAGTTTAAAGAAAAA

BH04TCATAAT.....GGAAGGGTTTTTCTCATTATTTATACTTTTGCAGTTGATTCACGGAAATGTACTTCTTACGA

BH05TCATG.....GGCAAGGGTTTTTGTCTATATTCAGATAGGAGGAGATTATATGAAAGCGGAGTTTACAAA

OI01TACTTT.....GGAAGGGTTTTTTTGAATTTTAGAATAGGAGGAGATTATATGAAAGCGGAGTTTACAAA

OI02TACTGT.....AGAAGGGTTTTTAACTTTCCCTATCTCTGTACAGAACTGTCAATATGCTAGTTTCATA

OI03 CTTCATTCTTATCATGTTTTT.....AATAGAAGGATGGATTATATATGAGCATACGGAAATGAAGATGAACCGA

OI04GAAATGTTTTTTTGTGATGGATGTTGTTTTATGATGATGATGTTGGAATAATCGAGAAA

OI05GTT.....ATGAAGGAAGTTTTTTTTTGTATCTATAATTTAGTACCGCGTTTTTATGACGAGGT

OI06 TA.....TCTCTT.....ACAGAAAGGGTTTTTATTTGACGAATATGCGGAAACTATACGAAATGGTTGCTGGAGAG

OI07TATTAC.....GGAATAGAGTTTTTTTGTTCATAGAAATGCTCTATAATTTGTTGGGTAAAGAAAAA

OI10TCTT.....TATGAAAGGGTTTTTTTAAATTAATAACGATAGATATGCGGATGAAATGCAAGTATGGTTT

OI08TCAAC.....GGAAGATTTTTTCTTGTGTCAGTGAGGGTTTGGAAAAATAAAGTGGGAACAGTTTGA

OI09TTAT.....GAGGGGGTTTTTTTAAATTAATACTTCTTATTACTTTAAAAATAATAAATTCGAAGGAGAA

OI10TCTT.....TATGAAAGGGTTTTTTTAAATTAATACTTCTTATTACTTTAAAAATAATAAATTCGAAGTATGGTTT

OI11 TCCCGCAGCGTCTTA.....TTAT.....GAAGGGATGCGGATTTTCTTATATTAATTTTAAAGGAGATGGTAAATGAAACAAACAT

O112 GTACGAA—TGCCTCTTCTCTG—TCCTT—EAGAGAAAGAGGGA—TTTTTATTTCCGCTGGGGTTGAGACATG—ATTGAATTCAGAAATGTAACA
 O113 TGTTA—TGCCTCTCGAT—TTCAC—ATCGGAGGCA—TTTTTAGTTTCCCGAAAAATTCACAACATGAGAAAAGAGGATTTA
 Ca01—GTGCTCTTATTT—TTTAAT—GGATAAGAAAGCAG—TTTTTATTTATTTATTAAGAGGAGAAATTTATGAGGAGAAATAGATTGTA
 Ca02—GCGCGAGG—TTAATT—CCAAGGCTTTTATTTATTCCTATTTTAAAGGGGCTAACTTTATGAAATCTCTCACTA
 Ca03 TAATAGA—TCTCTCTCT—TTAA—AAGAAGAGGA—TTTTTATTTGTTAAATAAGAACCAACTTATTAATTTTGGTTTATTTCTA
 Ca04—GAGCCAGAG—TTTAAT—CTCTGCTCTTATTTATTTTAACTAATGGGAAAAGGTGAATGACATGATAGAAAATAAA
 Ca05 AATAA—TCCGTTTCTG—TTAAT—GGAAATGGATTTTATTTATATATATTTTAGGTTAGATATTGAAGTTAAAAATAATAAA
 Ca06—TCCCTCTC—AACTAA—GAAGAGATTTAATTTTATATTTTATGTTAGAGGAAAGTGGATATAATGAGAAAGTTATTTAC
 Ca07—CTCTCTCTC—TTAAT—GAGGAGAGTTTATTTTATATTTGTTAGGAGGAAAGTGGATATAATGAGAAAGTTATTTAC
 Cp01—TCTCTCTCTAAATCT—TTAA—AGGTTGAGAGAGAGATTTTATTTAACAATAATTTTAAAGCCGCAATTAATAAAGTTTGT
 Cp02—CTCTCTCT—CTAG—GATAGGAGTTTATTTATTTGTTAGATAAAGGATAGATTTTATTAATGGATTAGGAGGA
 Lm01—TGCCTCGATCTG—ACCAAAAA—CAGAGCAAGCGTTATTTTATGCGCTTAAAGAGGGAGTTTGTGTTACATGAGAAATTTT
 Lm02—ACTCTCTAT—CTAA—AATAGAGAAATTTTATTTGCTTCTCATGAATAAATCTGATAAATCAGACAACTACTAGG
 Lm03—CCTCTCTGGGG—CTATATA—GTGCTAGGAGGTTTGTGATGAAATTTACTGATAAATACATAAATCAATCAAGAGGAGTGGATTTTA
 Lm04—CTTCTCTTAAAG—CTGT—CTTTAAGGGAAGTTTATTTATACATAAATAAATAAGAAATGAGGCGAAGAAATGAAACCAAG
 Lm05—AGCCTTCTCTATTCGCGCGCTTT—GTGCAAAATAGAGAGGCTTTTATATGAGAGCTATTTGGAGAGAAATGAGGAGGAAAAATAAAATTTGG
 Lm06 GTATCGAAACTTCCAAA—CCTCTCTAG—TTCT—CTAGGGAAGTTTATTTGTCATAATCAGAAATGTAGAAAAGGTGATAGAGATGATTACGTACA
 Lm07—GCCCTTCTCT—CATC—AGCAGAAAGGCTTTTGTATATCAGAAATGTAGAAAAGGTGATAGAGATGATTACGTACA
 Li01—CGCTTCGATCTG—ACCAAAAA—CAGAGGAAAGGCTTATTTTAGCGCTTAAAGAGGGAGTTTGTGTTACATGAGAAATTTTA
 Li02—CTTCTCTAT—CTA—AATAGAGAAATTTTATTTGCTTTCATGAATAAATCTGGAATAATACTCAACATACTAGG
 Li03—CCTCTCTGGGG—CTTATATA—CTGCTAGGAGGTTTGTGATGGAATTTACTGATAAATACATAAATAAGAGAGGTGGATTTTA
 Li04—CTTCTCTTAAAG—CTGT—CTTTGGGGAAGATTTTGTGACATAAATAAATACTAGAAATGAGCGAAGAAATGAAATCAAG
 Li05—GCTTCTCTATTCGTCGCG—TT—GAGCAAAATAGAGAGGCTTTTATGAGAGCTATTTGGAGAGAACTAAAGGAGGAAAAATAAAATTTGG
 Li06 GTATCGAAACTTCCAAA—CCTCTCTAG—TTCT—CTAGGGAAGTTTATTTGTCATAAATAAGGAGGATAAGGTGATAGGTTGGTAAAGGC
 Li07—GCCCTTCTCT—CAAT—AGCAGGAAGGCTTTTGTATATCAGAAATGTAGAAAAGGTGATAGAGATGATTACGTACAG
 Ss01—CTCTCTCT—TCAAT—AGGCTGAGAAGTTTGTGTTAAGGAGGAAAGAACAAATGACAAAATTACAGATAGTACTT
 Ss02—TGGTCTCTGAATTT—TCTATC—GATATGAGAAAGCA—TTTTTATTTAAGCAACACAGGAGGAAATCAACCTGATTTGAATTA
 Ss03—GGCTCTCTCTAT—ATTA—ATAGTTAGAGGCTTTTATTTAGCTCACAGAGAGAGAAATTTTGTATAATAAATTTAAAGGA
 Ss04—AGCTCTCTCTCA—TCGTT—TGTAGAAAGAGGATTTTAAATGGAAAGCAGTAAAGGATGGAAAGTACATAAAAAAGAGCA
 Ss01 AGACTCCGAAAACCTCCACGAACC—CCACCGACCGCGCTCGACCTCGGCCCGCCCGCTGAGCTGCCCGG

[illegible]

Ba15	GCAAAATTAAA	CGCTTCCTCT	ATATAA	AGTAGGAAAGGTTTTCTGTATGCTTGTGTGGAGAAATAAAATGATGTGCGCAATCTGTGGCA
Ba16	ATAAA	CGCTTCCTCT	CTCTA	AAGCGGAAGGTTTTTGTGTTTGAAATGTGGAGGACATTCAAAATAATAAAAGTAATGAGA
Ba17		CTGCTCTCTCTG	TTTT	CAGGAAAGGGCAGTTTTTATTTTGTATAAAAGAAAGGAGAAATGAGAAATGAGAAATCATGCGG
Bc01		GTGCTCTCTCTCTATC	TTTATGGTT	GATAAGAAAGGAGAGACAGTTTTTATTTTACCTCGAGAGCTCTGCTTCAAAGTTTTTCACAGCATATAGGAG
Bc02		GCCTACTCTAT	TTCC	ATAAGAGTAGCTTTTTTTATGGCTAAAGTTAAAGGGGAAATAGGTAGTGGAGTATGTTTTT
Bc03		TCCTTTC	GAATG	GGAAAGATTTTTTTTATGAAATAAAAGGGGGCTTTCGGCTGAGCGGTACGGGAACAT
Bc04		GACCTCTCT	CGAT	GGAAAGAGGTTTGTATTCAATTAGAAAAAGGTTGAAACTAGGGAGAGATGGTACTTTGA
Bc05		TGAGCTCT	ATGT	AAGAGTCTATTTTTTGTGTATAGAAAGGAGTGTGCATGTCATAATTCATTTTTCAAAAATA
Bc06	AA	GGCTCTCTCTA	TCTTT	TAGAAAGAGGCTTTTTTATGTGAAATAATAAGGGGGAAGAAAAATGGAGCGCAGCAGGAGTAA
Bc07		GTCTCTCTAT	ATGT	GTAGGAAGAGTTTTTATTTAGAGAGGGGGATAGAGTGAAGTTTGTATGTAACGTATTTTT
Bc08		CTCTCTC	TTATC	GGAGAGGTTTTTTATTTGCAAAAAAACCGATTACGAAAAATTTATATAAGAAAGAAAAAGG
Bc09		GGCTCTC	CTTAGCT	GACAGGGTTTTTTATTTAACTAGGAGGTATAACAATGAGCGGAATTAAGCGACAT
Bc10		GGCCCTCTCA	AAG	TGAAGAGGGGGTTTTTATATTGATAGAAATGAGGGAGATTTGTGAAAATTACTAGATTTATTA
Bc11		GCCTCTC	TCTT	AGAAGGTTTTTTATTTTATATCAACTAATGTTTCAATTTAAAAAGGAGGAATTTTC
Bc12		CCCTCTCT	TAGT	GGAAAGAGGTTTTTCTACGTCAGAAAAAACCTCTGAAATAAAAAAGGGGGAGAAACGAT
Bc13	GCAAAATTAAA	CGCTTCCT	CTCTATTATGT	AGGAAAGGTTTTTCTGTATGCTTGTGTGGGAGAAATAATGTAATGTGCGCAATCTGTGGCA
Bc14	ATAAA	CGCTTCCTCT	CTCTA	AGGAGAAAGGTTTTTTTGTGTTGAAATCTGGAGGACATTCAAAATAATAAAGTAGTGATA
Bc15		CTGCTCTCTCTCTG	TTTT	CAGGAAAGGGCGAGTTTTTATTTTCTATAAAAGAAAGGAGAAATAAGAGATGGGAGAAATCATGGG
Bc16	TAACCTGTACGTAAAAA	TCCCTCTCTC	TCAATAT	GAAGAGGGATTTTTTATTTTTCATTTCCCTCATCATCATCCAACTTAATTAATTAGGAG
Bc17		GCAGCTCTG	TTATTTT	GAGAGGTGTTTTTATTTTGGAACGTATATTTAAGGGGGAATTAAGATGAAGAAAGTAT
Bc18		GCCTGATCTTA	TTTTT	TAAAGATCAAGGGTTTTTGTATTTCTAAAAAGAGAAAAAGGGAGTAAATGGAAAAAGTACGTTTCA

Cons

Bs01 **G**GGACTTTTAGAAGATTTTGC AAAAGACAGGTTGTAATCGGTGACGGCGCCATGGGGACGCTCCTCTACTCCTATGGCATTTGACAGGTGTTTTTGAGGAGCTCAATATTTTCAAAAGCCGGAGGA
Bs02 CGAAAATTAGCTCAAAATTGGGAACCGTAGCGATGAAGTCAAGGGAACAGTAGTGCTCCTATCTATTTATCAACAGCATACCGCCACAGAGGATCGAGAAATCTACCGGATTTTGATTTATG
Bs03 GTGTTTTACATATAAGGAGGAGAAACA**ATG**ACAAACCATCAAAAACATCGAAATTTAGGAATTCGAGAAATCGACCTGAACCGGGAATGGA AAAAGCACTTTGAAGCGATTTGGAAGGCGAG
Bs04 GGAGTCACAAAACACCTTTATACGAAACGTTAAATGAAAGCTCCGCTGTGGGTTGGCGTGAAGCTTGGCTATTTCCAGCAAAAGCACGCTGACATGCCAGGAGATCGGAGACGGC
Bs05 GCGACATATCTCTGACCGAACCGGAGCCGATACAGAGAAAGAAAGCAAAATCGCAACAGGAATGACAGTAGGCTCTGGACTGATCTGCCCTTTGTAAACAGGAGCAAAATGCAA
Bs06 CACATTACATGGGAAAGACAAACCGGACAGAAAGAGCTTGAAAACATTTGCTCCGAAAGGAGAAAGAA**ATG**TTAACGTATGATAATTGGGAAGAACCAACGATTACATTTCCGGAAAGA
Bs07 TGGTTACAGCACCGTACAAACGAAGAGGACGAAAGAGCTTGAAAACATTTGCTCCGAAAGGAGAAAGAA**ATG**TTAACGTATGATAATTGGGAAGAACCAACGATTACATTTCCGGAAAGA
Bs08 TTGTATAAGAGGAGCGAGTTCCGTATCATATATACAAAGTCTTCCGGAGGCCCTTGTGCAGGAGGAAGCAAAATCG**ATG**AGTAAAAATCGTCTGTTTATTTACATCAGAAATCTGTTACGGAG
Bs09 GTGATTGCT**ATG**ATCAATCTTCAGGATGTTTCAAAAGTTTACAAGTCTTCCGGAGGCCCTTGTGCAGGAGGAAGCAAAATCG**ATG**AGTAAAAATCGTCTGTTTATTTACATCAGAAATCTGTTACGGAG
Bs10 ACGAAATGTTGCAGGACAAAAACAGAAAAACAAACGCAACACCTTTCGCGCCGATCATGTGGGACGCTTGTCTCCGTTAAAGGAAGGCCCGGCAAAAAAGCGGCTGGT
Bs11 ACACCCGACAGACAAAAATCACTTCAAGAAAAAAACCGCCGTTTCGCGCGGATCAAGTCGGAAAGCCCTGAAGATCTGAGCCCGTCAAAAAGCGCGGCTCAAAAAGCGCGGCGGCGAA
Bh01 CCGAAAAACAGCAAGACAAAAAAGAAACAACTTGAATGAGGAGGCCGTTG**ATG**AAAAAAATTTACGTAAATCCACGAAACAGATGAATGGACGGAGTTAAGGAAGAAAGTTACTCTCGAAAT
Bh02 AAGGGAAATTAATCAG**ATG**ATTAAAGTTGGTGATCGGAATTTGCCACCGTTGGCAAGGTTGTGTCGAGAGCTAGTTCAAATGGAGCGAGGATTAAGGAAGAAAGTTACTCTCGAAAT
Bh03 TGAAGGCTTTCGCAACTTGGCAGCGAGCTGATTTTCCAAATAGATGGATAGGAGGAGCAAC**ATG**AAATCGTAAAGAAATTAAGAAACAGCTTTAGTACAAATCGGAAATCGAAATGGATGATC
Bh04 ATCATGTTTGGCAGCCCTGGTTGGTAAGGGTGCTTAAGGGAGGATATTCGTAA**ATG**GCAGATACAAAGAGTCTGCTGTTTATTAATCAGAGTCTGTTACAGAGGACATCCTGA
Bh05 TAATGATTTCGCTCTTGAGACGAAATTTCCGAGAGTGAGAGTTTTCCTCTCTGACTGACTTTTCGTTAAATTTGGTAACGGGTAGACGAACTGATATATTTTAAAGAAAGAGGGCTT
O101 GACGTTTGCAGGAAGGACCGGTAATAGCAGGAGAGGGGTACTTACAGGACGTTTCGTTTGTACCAGAGTAGCCCTTGAAAAATCCGGATGCGT
O102 GAGCAAGACCTACTCTATAAGACTAGCCCAATCTAAAGGAGAAAGAGAAATTAAC**ATG**ACAAAAACAGTTATTAAGCACATGAGGAGTTGGAAATACGAATACTTTAAATACAAAT
O103 ACAAGAAGAAATGATCTAAATTGAGAAATTAATTGCACTTAATCAATTTTAAAGGAGAAAGAGAAATTAAC**ATG**ACAAAAACAGTTATTAAGCACATGAGGAGTTGGAAATACGAATACTTTAAATACAAAT
O104 CACTATGAATCTAAAGAGAAAGAGTGATAGGGGTAGATCAAGTTTCCCTTGATATCAAAAAGGAGAAATATATGGCATCTGTTGGATATAGGGTGCAGGTAAAGTACGCTTTTACGT
O105 TCTTTTATGGCACTTTGAATAGGATAGAAAGTTATAAGAGATCCGTACCAACATATATCAAAAGGAGAGTTTAGCCCTT**ATG**GGCTGCAAAATCGACGTTTATTTACTTCAGAGTCAGTAACCT
O106 TAAGAGGAGGAATAAGAA**ATG**ATATCCATCGAAGGTTAAGTAAAGTATTTTCATTAAATAAAAAAGACATCAAAAGCTGTAGACTCATTTGACCTCAATATTTGAAAAATGGCGATATTTAT
O107 TAAAAAACACGCAATCTCCTATTTTGTATTGTTTAAACCACTAAACCAAAAGGAGATCGGTGCA**ATG**AAATTTCTAACATAACATTTACCTGGGTTGGAAAGAGAAATATA
O110 CTGGTTGCCGATTTTGGAGGGTGGTTGCGTAAATGTAGAAAGATGAACAGATGCCCTCCTACTTTTGAATATGCAAAACAGGTAAATTCAGCACGCGGAAAGAAATGGGGATATGATACGACTTT
O108 CTTCAAAATATGAGTAAACCAATCAGGTAACATAAGTAAAGTAAAGGAGTAAAGTCTGTAAGTTTATAAAATCTAAAAATGAAGAGGAGAGCGGTGTA**ATG**CCAACTATAAAAAAC
O109 ACAC**ATG**CTCTAAATTTCAATCTTTGCAAGCAGAAACAACTTACTTCTATGAGGAGACAGGAACCAAGCCCATCAACTGGTTCAGTGCTCCAAATTTATCAACTACGCTCTATGTGT
O110 CTGGTTGCCGATTTTGGAGGGTGGTTGCGTAAATGTAGAAAGATGAACAGATGCCCTCCTACTTTTGAATATGCAAAACAGGTAAATTCAGCACGCGGAAAGAAATGGGGATATGATACGACTTT
O111 TGTGACATTTCCGGCAGCCCTCCGAACTATCTAGATCTGAAATAAGTACTACATTTTAGGGAAATCAATTAAGTGAACAGAAATCTTATGTGACCCATATTTCTGTTAAAGATGTACC

Oi12 AAGACATTCACACTAGGAAAAAGAAAGTAAGAGCTGTTAAAGAAAGTATCTCTAACGATCGAAAAAGAGAGATATTATGGAAATTATGGGTTACGGGTGCGAGGAAAAAGTACCTTGCTT
Oi13 **IG**TCCACATCGATTGTAAAAAGGAGCTCCGGGTCAATTACGGATTGGCGGATGCTTGGAGGAATTCCTGTACTGCTTGAAAGAACTGTAGTTAATCGTATACAAGTTATCCGAGGGA
Ca01 GAAATTTTGGAGACAAAAGCAGTTCAATGGGGAGAGTGGTTTGGAGACGAACCTGGGGCAATAAGCTACCCCAATAACCAAAAGTTTACCTTTAGACATGAAGGCTTAAATAAGGAATG
Ca02 AAGAAATTTGTTAAATAACAAAAATTTTAGTTTATAGATGGTCTATGGGAACATGTATCAATCCCTTAACTCTAGATGAAGGCGACTTTAAAGGTTCTTATCTTTGACATGTCAATTCCTT
Ca03 TTAAGAGTGGTGTATAGGACATATTTTATTAAGAGAAAGAGAGAAATACCTCCCAATTTCTCCCTTCAATTCCTATAGATTTTACCCAACTTATCCCAATCTATCTCTAAAAATATTTTACTA
Ca04 AAATGTTTCTAAATATTTTTCAGGAAATAAGGTTCTTAAAGATTTGATCTGAAGATTAAAGCGGAGAAATAATTTGGAAATTTGGTCTATAGTGAGCTGGAAAGTCAACATTTACTTAG
Ca05 AAAATACCTGATAATCTCCAGCAGCAAAAACCTTTAAATGAAGAAAAATATATTTTATGGATGAGGATAGAGCCTTATCAACAAGATATAAGACCTTTAAATAATTTGTTATAGTTAAACCTT
Ca06 AAGGGGATTTAAAG**AG**AGTGAGGAAAGAAAAATTTGGTTTGAACATTAAGTTCTACAGGATTCAGGCAAGTTGCTGATCCCACTACAGGATCAAGGATCAAGAGCTGTACCTATTTATCAACAACA
Ca07 ATCTGAATCAGTAACAGAGGGCATCCAGATAAAATCTGGATCAAAATATCAGAGCTATTTAGATGCCATATTGGAAAAAGATCCCAATTTGGAAGAGTTGCTTGTGAAAGTACAGTGAC
Cp01 TAAATTAAAGCTTTAAAGATAATTTTACGATTTATACCATTAATAATTTTGGTATCGAATTTATTTGGAACTTTAGTAAATGAAGGTTGGAAATAATAAGGTTGGAAATTTAGCTATCTTATTTAGTTCCAGGA
Cp02 GAGAAA**AG**AAAAAGGAAAGTTTACGATTTATTAAGTTATTTTGGTATCGAATTTTGGAACTTTAGGATTTAGCAGCATGTGGAGGAGGAAAGGTTGAGGAAAGGAGTTGCTTGTGAAAGTACAGTGAC
Lm01 TATTAGTAGCGGTATCTCGGTTTTCGCTTGGTTTACGGCTTGGGAGGTTCTGGGCTAGTTTCAGACAAAAGCAACCGTTACGCAATTTAGGATTTAGCAGCATGTGGAGGAGGATTAAGCTTAAGCACAGAAAAA
Lm02 GAGGAAAAAG**AG**AAAAAAATTAACAAAGGGTTAGGAAATTTTACTTCATCAAGCCTTGTGTAGGATTTAGCAGCATGTGGAGGAGGATTAAGCTTAAGCACAGAAAAA
Lm03 **IG**AGTAATGAGTATAAATTCGAAACAATTCAGATACACCGCGGACACACACCGGACGGAGATACACATTTAGAGCGCTACCTATTTATCAAAACGACGTATACACATTTGATAGCCCGG
Lm04 TAGCTCCATTTATGCAGATCATGTGGGAAGTATTTACGGCAAAAAGGAAATTAAGAGCGCAGAGAGAAATTCCAAAGTGGCGAAATAACAGCCCTTAGAGTTGGCGAAAAATCGAAAAATA
Lm05 CTAAGAACCGTCATCTATTACATCAGAAATCGGTTCTGTATGCAGATCCAGATAAAATTCGAGATCAAAATCTGTGAAATCTGGAAATGGTGGCAATTCGGAAGAGGAAATTTGGTGGCTGAAACGAAAGGCTCT
Lm06 GATTAGTTCAAACCTTGGGTATCCGAGACTTGGGAGAAACGTGAATGGAAACGTGGTTAGAAATTCGAAATTTGAATTTGAACAAAGCGGAGATTTTCGGAGTTGAGGTTATTCGGAGAGCTGGTAAAAAG
Lm07 AAACGTTGTAAAGAAATACACGTCCGAAACAAAGTTCTCGCAGTCGATCATGTGCAATTTAGAAATTTGAACAAAGCGGAGATTTTCGGAGTTGAGGTTATTCGGAGAGCTGGTAAAAAG
Li01 TTAGTAGCGGTATCTCGGTTTTCGCTTGGTTTACCGGCTTGGGAGGCTCTGGCGCTAGTTTCAGACAAAAGCAACCGTTACGCAATTTAGGTTAGCAGCATGTGGAGGCGGAAAGTGACGATAAGCCCTTAAGCACAAAAAGAA
Li02 GAGGAAAAAAAG**AG**AGAAAAATTAACAAAGGGTTAGGAAATTTTACTTGCATCAAGCCTTATTTAGGTTAGCAGCATGTGGAGGCGGAAATTTATCAAAACACATCGTATACATTTGATAGCCAG
Li03 **IG**AGTAAATGAGTATAAATTCGAAACAATTCAGATACACCGCGGACATACACCGGACGGAGATACGCAATTTAGAGCGCTACCAATTTTCAAACTCAAGAAATTCAGTGAATTTGAAAAATG
Li04 TGGCACCAATTTATGCAGATCATGTGGAAAGTATTTACGGGACAAAAGGCAATTAAGAGGCGCAGGAGAAATTCCAAAGTGGCGGAAATTTAGATGCAATTTTTCAAAAGATCCGGACCGCACGTG
Li05 CTAAAAACCGTCATCTATTACATCGGAATCGGTTTCTGTATGGACATCCAGATAAAATTTGCAGATAAATTTGCAGATCAAAATCTGTGCAATTTTAGATGCAATTTTTCAGAGGAGGAAATTTATGGCGGAAACAAAAAGCTCT
Li06 GATTAGTTCAAACCTTGGGTATCCGAGACTTGGGAGAAACGTGAATGGAAACGTGGCTAGAAAAGTTTGGAAATGGTGGCAATTTTCGGAGATTTTCGGAGTTAGTTGTCATGCAATTAACCTGGCA
Li07 AACGTCTGTAAGAAATATACGTCCAGAAATAACAAAGTTCTCGCAGTCGACCATGTGATTTAGAAATTTGAACAAAGGTGAGATTTTCGGAGATTTTCGGAGGCTGGTAAAAAGT
Sa01 TAAATCTAGGGAATTTATACAGAAATCTGGGAAGTCATAGATAAATCTGGTTTGAAGATGTGGTATTCATGGACAACCAATTTAGTTGATTTTTCGGAGCTATTTGTTTCGGAGCAGGAAAAAG
Sa02 AGAAGTTGTTAAAGAAATCTCGGACTAAAAATAAGAGTCTTGTGTAGATACACGTTAAATTTATCGATTTCGAGGAGTGAATTTATGGCGTCAATTTGGTTTTCGGAGCAGGAAAAAG
Sa03 GCAAACT**AG**TTAAATAACAAACGATTTATCTCAGAGTCTGTACAGAGGACCCAGATAAAATCGCTGACCAAGTGTGAGATGCAATTTAGATGCTATTTTAAAGAGGACCC
Sa04 ATGCTTGGGCATTTATCCCTTGTATTTATTTTGGCGTTGTTTATAGCGGTAGGTTATTCACAGGTGATTTTACTTCAATGCCCCATTAATGTTGCAATTTACGATAACGGTAAATTTGTGG
Se01 AGTGGCGGCACAGGTTCCGCTCGGACCGGCTCTCGCCTGCGAAAGTGTGTTTCGGCCCCCTCGAGATCGCCTACGACTTCTCGGACTACGACGCCGAGAGCTGCGCAAGCGGATCGGAAG

Cons

C01 GCCGGGACGAGCCTGATCCGGGTGACCGGATCGGGCGCCGTGCCCAATTACAGACACCTCTCTAGTGTTCGAGAACGCGGAGCACGGCGCTGACCTGTTCGGCGCTTCGCAAGCGGGGC
 T01 AGTCAGTCACTGAAGGGCATCCTGACAAGA TCTGTGACCAAGATTTCAGATGCCAATTTCAGATGAAATTTTAAAAAAGACCCCTTACGCCCGGTGGCATGTGAGACAGCTGTAACTACCG
 T02 GTGTAATAACCTTAAGAAAGAAATTTGTATAAAGTGTGAAATTTCTCCCCCAAGAGGATAGATGTAACATAAATCGAGGAAGCTCGAAAACTTAAAGGTGTGGCAGATGCTCT
 T03 ACTGTTGGGACAGGAGTATTAAATAGTTAATCTAGAGGGAGATATATCAAGGAGAGTACGGATTTTATCCGGAGATAAAGAAAGTGTGTGAAAGGATTTGCACAAAAAGAGAAAA
 F01 GTTTACATACTTTACATCAGAAATTTGTTTACCAGGACATCCAGATAAAATTTTCAGATCAAAATTCAGATGCAATTTTATGATGTTTAAAGATGACCCCTAAATTCAGAGTTGCCCTG
 F02 CACTTGAAAAATGTAAATAAAATTTTATCCCAATAAATTTGATGCTGTAAAGATGTTAAATTTAAAAAGTTAATGAAGGAGATATCTTTGGAAATTTATAGTTTAAAGTGTGCTGGAAAAATCTT
 D01 ACCGTCTCCCCCGCCCTTGCACTTCGAAGCGTCAAGCAAAACCTACCCCGGCCAGCCGGCGCTGAGCGATTGACCCCTCACCGTTGCGCGGGCAGCCGCAACCCGCAATCAT
 D02 CAGCGCATCCTGATTCGAGCGCGCCCTGGGTACGCGACTTCAGCGAGCCAACTCACCGAAGCGGACTTCCGCTGGGACGAAAGCCGACCCCAACCGGATGTACCCGGGGCAACTTTCGAC
 X01 CCAACGATTTGTTGACACCCCGCCAGCAGCGACGGCATCACTGCGTGGCGGGGAACTTGTCTATCGCCCTGCCGATGCCGCAATGCCGAGCTGCGGCTGCGCTATG
 X01 CCACCGCTGACACCTACACGCCCGCGCTGATAGCGACGCCCGCGCTGCGTGGCGGAGCTGCTCATCAATCTACCGATGCGCCACGCCGGCCAAACGCGAGCTGCGGCTGGGCTACG
 S01 CATTAGAACTAGGTGAGTTTAAAACTGAATCTGTTGTAACGATTTGATCTACGTACGTTATGAACATGTAGGACTTCCTGCTCAACCCCTTGTGCTGTTTGGCATGCACTTACTG
 S02
 G01 CATGCTCGAAGAAACAATCCGTCACTTCGAAACGGATCTCAGGCTGGAAGCGGCGGATACTGCGGCCCATACCTTGGCTACGAGACCTACGGCCGCTGAACGCCGACCGGTCCAA
 G02 GGTCTCGACTGGGATACCCGACCGGGCGGTGACGGTACCCATCTACAGACGGCAACCTTCGGGCAATCCGGGATTTGGCCAGAGCACGGGCTACGATTATCCCGCTCCGGCAACCC
 Ba01 ATTAAGCATTTGTAAGCGGACGGTACTATTATAGCGCATGTAGCGCTGGTTCCGATAAAGAAAGTAAAGCGTTAGATGAGAAAAAGATTACTGTGCTGTAAACAGCGCGGGCCGATGA
 Ba02 AAACAAAGTAAATTCATGTGTTTGAAGGGTTATGGAAGTGTATGTAATTAATAAATTCGGTTATGCTGTTACACTATGCGATTACGACAGTGACATTTCTAAATTAATGCTATTAAT
 Ba03 GAAATCAAAATGAAAAAGAAAGTTGTACCCGGTATGTCATCAGTTGTAGGAGTAAAGTATTTAATACTGTTGCGGTAGTTATAAAACGAAGCAAGCGGAGCAAAATGCAAAAGACGAG
 Ba04 GGGGAAAAATGATTTCTTTTAAATGTAAGTAAAGTATGAAATCAGTTGGCAATCTGTTGCGGTGGAGGATGTAACGTTATCAGTTGAGAAAAAGCGGAAATTTTGGCAITATCG
 Ba05 GACCTATCACAACATTCGAAAAATAATATGCTCAATTTCCCTGGAACACCAAAAAATCAATTTAGAAAGCCATTAACAGCGTTGAAGAACAGGTTATCAAGTTACAGATTTCCATTTCTG
 Ba06 TTGAGGAAGTGTCTGAGAGAAATTCAGCGGATGCTTCTGTATGAAATATGTTTCAATTTACAAITTTAGTTCTGAAATTTACAGGTGATTCTTAATAATTGCCGATACTTGTAGCAATTTCA
 Ba07 AAAGAAACGAGAGGAAATGGTTTGGCTTTTACCCTTGGGATAATTTTGGCGCTATTTATAGTTCTGAAATTTAGTTCTGAAATTTAGTTCTGAAATTTAGTTCTGAAATTTAGTTCTGAA
 Ba08 AATATAGAGTAAATAAAGTTGACTATTAGAGAGGGGAAATTAATAGAACAGATTATCAACAAAATAGTAGTAGCAATTCGAAATTCGGAAATTTGATCAGCAITTAACGGGATATAGGACTTTGGG
 Ba09 CGTCACAAAGAAAAACAAITGGAAGAGATTCGAAAGAAATAGGAATAAGTAACATAGAAACAAAGTCAATGATAATTCATGCGAAATTCGGAGATTGTTAAATCAAGAAATTTTACGCAITCTAGAA
 Ba10 GGGGTTGGGAGTACTGTGACACTCGAAAAATACGTAAACACTGCGTAGTACAGTTTATGAATATATGATAGAGCAAGATAAGCCAAATATCATTTGTTAGATATTCAGAAACATATCTGTTTC
 Ba11 GGATATTATTCATTACAGAAAGTAAACCGCTGTACAAATATGCGAAAGAACATGTTTATTTTGAAGAAAGAAAGCAATGTAGTTTGTGATGAAATTTGGAGATGGAATTTAAATTTATGTTTC
 Ba12 ATTTAATCCATGATGATTCACATAAATTTAGAAAAAAGCTGAGCAAAATTCGACTCGGTTTAAACAAITTTGGCTCTTGGACTCAITTTGCCACACTTATTCGAAAGACAGTFAAGCAGCAT
 Ba13 TACATGTCAACTATCGAAACAAAACTAGCCCAATTCGAAACCGGAGTGAAACTACAAACGAACTGTTAATCCCGCTGTTTACTTTTCACTGCTTATCGTCAAGGAATTTGGTAA

B. Cobalamin Riboswitch Alignment.

ID	Start	End	Accession	Organism
Atu01 +	70441	70625	NC_003304.1	Agrobacterium tumefaciens
Atu02 -	441331	441136	NC_003305.1	Agrobacterium tumefaciens
Atu03 +	877645	877833	NC_003304.1	Agrobacterium tumefaciens
Atu04 +	921717	921886	NC_003305.1	Agrobacterium tumefaciens
Atu05 -	1640563	1640420	NC_003304.1	Agrobacterium tumefaciens
Atu06 -	2810076	2809899	NC_003304.1	Agrobacterium tumefaciens
Bha01 -	466904	466746	NC_002570.1	Bacillus halodurans
Bha02 +	528894	529051	NC_002570.1	Bacillus halodurans
Bha03 +	870599	870748	NC_002570.1	Bacillus halodurans
Bha04 +	1661078	1661219	NC_002570.1	Bacillus halodurans
Bsu01 -	3403719	3403620	NC_000964.1	Bacillus subtilis
Bja01 +	2232813	2232975	NC_004463.1	Bradyrhizobium japonicum
Bja02 +	3617311	3617490	NC_004463.1	Bradyrhizobium japonicum
Bja03 +	3630677	3630857	NC_004463.1	Bradyrhizobium japonicum
Bja04 +	3634122	3634284	NC_004463.1	Bradyrhizobium japonicum
Bja05 -	5279669	5279495	NC_004463.1	Bradyrhizobium japonicum
Bme01 +	679030	679218	NC_003317.1	Brucella melitensis chromosome I
Bme02 +	717388	717585	NC_003317.1	Brucella melitensis chromosome I

Bme03 + 559758 559950 NC_003318.1 *Brucella melitensis* chromosome II
Bme04 - 973106 972933 NC_003318.1 *Brucella melitensis* chromosome II
Ccr01 + 502968 503156 NC_002696.2 *Caulobacter crescentus*
Ccr02 + 1925017 1925166 NC_002696.2 *Caulobacter crescentus*
Cte01 - 409802 409630 NC_002932.3 *Chlorobium tepidum*
Cte02 + 422045 422244 NC_002932.3 *Chlorobium tepidum*
Cte03 + 443769 443951 NC_002932.3 *Chlorobium tepidum*
Cte04 + 584183 584411 NC_002932.3 *Chlorobium tepidum*
Cte05 + 882576 882770 NC_002932.3 *Chlorobium tepidum*
Cac01 + 1509969 1510116 NC_003030.1 *Clostridium acetobutylicum*
Cac02 + 2557903 2558041 NC_003030.1 *Clostridium acetobutylicum*
Cpe01 + 248269 248429 NC_003366.1 *Clostridium perfringens*
Cpe02 + 1241749 1241887 NC_003366.1 *Clostridium perfringens*
Cpe03 - 1431291 1431152 NC_003366.1 *Clostridium perfringens*
Cpe04 - 1549481 1549348 NC_003366.1 *Clostridium perfringens*
Eco01 + 4160983 4161133 NC_000913.1 *Escherichia coli*
Fnu01 + 934517 934658 NC_003454.1 *Fusobacterium nucleatum*
Lig01 + 1347854 1347994 NC_004342.1 *Leptospira interrogans* chromosome I
Lmo01 + 1179829 1179979 NC_003210.1 *Listeria monocytogenes*
Mlo01 - 1101076 1100918 NC_002678.1 *Mesorhizobium loti*

Mlo02 + 1149143	1149308	NC_002678.1	Mesorhizobium loti
Mlo03 - 4044577	4044416	NC_002678.1	Mesorhizobium loti
Mlo04 - 4957334	4957164	NC_002678.1	Mesorhizobium loti
Mlo05 - 6170855	6170715	NC_002678.1	Mesorhizobium loti
Mlo06 + 6749148	6749315	NC_002678.1	Mesorhizobium loti
Mle01 - 1130394	1130222	NC_002677.1	Mycobacterium leprae
Mfu01 - 309822	309703	NC_000962.1	Mycobacterium tuberculosis
Mfu02 - 1261701	1261497	NC_000962.1	Mycobacterium tuberculosis
Pae01 + 1381520	1381688	NC_002516.1	Pseudomonas aeruginosa
Pae02 - 3261415	3261204	NC_002516.1	Pseudomonas aeruginosa
Pae03 + 3265563	3265728	NC_002516.1	Pseudomonas aeruginosa

ID	Start	End	Accession	Organism
Pae04	-	3305780	3305629	NC_002516.1 <i>Pseudomonas aeruginosa</i>
Ppu01	-	2765203	2765045	NC_002947.3 <i>Pseudomonas putida</i>
Ppu02	-	2768953	2768785	NC_002947.3 <i>Pseudomonas putida</i>
Ppu03	+	3857563	3857746	NC_002947.3 <i>Pseudomonas putida</i>
Ppu04	-	3981958	3981816	NC_002947.3 <i>Pseudomonas putida</i>
Rso01	-	2609233	2609017	NC_003295.1 <i>Ralstonia solanacearum</i>
Sme01	+	954780	954943	NC_003047.1 <i>Sinorhizobium meliloti</i>
Sme02	-	1999747	1999574	NC_003047.1 <i>Sinorhizobium meliloti</i>
Sme03	-	2122891	2122516	NC_003047.1 <i>Sinorhizobium meliloti</i>
Sme04	+	66265	66456	NC_003078.1 <i>Sinorhizobium meliloti</i> plasmid pSymB
Sme05	+	580403	580578	NC_003078.1 <i>Sinorhizobium meliloti</i> plasmid pSymB
Sco01	+	1037869	1038053	NC_003888.1 <i>Streptomyces coelicolor</i>
Sco02	+	1045899	1046106	NC_003888.1 <i>Streptomyces coelicolor</i>
Sco03	+	1051420	1051563	NC_003888.1 <i>Streptomyces coelicolor</i>
Sco04	-	5688395	5688291	NC_003888.1 <i>Streptomyces coelicolor</i>
Sco05	-	6532337	6532191	NC_003888.1 <i>Streptomyces coelicolor</i>
Sfi01	+	4183416	4183566	NC_004337.1 <i>Shigella flexneri</i> (*)
Son01	+	826836	827026	NC_004347.1 <i>Shewanella oneidensis</i>

Son02 + 1071692 1071874 NC_004347.1 *Shewanella oneidensis*
Sti01 - 2114053 2113918 NC_003197.1 *Salmonella typhimurium*
Sti02 + 4347871 4348024 NC_003197.1 *Salmonella typhimurium*
Tma01 - 84288 84144 NC_000853.1 *Thermotoga maritima*
Tte01 + 395153 395353 NC_003869.1 *Thermoanaerobacter tengcongensis*
Tte02 + 396075 396275 NC_003869.1 *Thermoanaerobacter tengcongensis* (*)
Vch01 + 145142 145306 NC_002505.1 *Vibrio cholerae* chromosome I
Vvu01 + 1165724 1165882 NC_004459.1 *Vibrio vulnificus* chromosome I
Xac01 - 3631166 3630987 NC_003902.1 *Xanthomonas campestris*
Xax01 - 3758428 3758245 NC_003919.1 *Xanthomonas citri*
Ype01 - 4393235 4393008 NC_003143.1 *Yersinia pestis*
Aca01 - 340 170 M34485.1 *Acinetobacter calcoaceticus*
Avi01 - 388 214 U45329.1 *Agrobacterium vitis*
Bff01 + 580 762 AY043208.1 *Bacteroides fragilis*
Bmg01 + 1211 1350 AJ000758.1 *Bacillus megaterium*
Lma01 - 76392 76234 AL499620.1 *Leishmania major*
Pff01 - 543 373 AY033236.1 *Propionibacterium freudenreichii*
Rca01 + 105327 105521 AF010496.1 *Rhodobacter capsulatus*
Rca02 + 116991 117174 AF010496.1 *Rhodobacter capsulatus*
Rca03 - 39849 39672 AF010496.1 *Rhodobacter capsulatus*

Rsp01 +	201	341	B07728.1	Rhodobacter sphaeroides
Sbi01 -	330	147	BH245584.1	Sorghum bicolor
Sgi01 -	9209	9035	AF263012.1	Streptomyces griseus
Svi01 -	1235	1052	U27616.3	Stealth virus 1
Zmo01 -	24942	24808	AF193754.1	Zymomonas mobilis
Zmo02 -	4323	4162	AF193754.1	Zymomonas mobilis

Accession numbers are for Genbank sequence entries. Start and end positions are the 5' and 3' nucleotides of the first interior UG base pair of stem P1 (orange). Secondary structure (SS) and sequence consensus (Cons) lines are shown above the alignment. In the consensus sequence, uppercase and lowercase letters represent $\geq 90\%$ and $\geq 80\%$ conservation at a position, respectively. The degenerate bases R (A,G) and Y (C,U) appear only when a single base is not 80% conserved. Sequences marked with an asterisk (*) were excluded when determining the sequence consensus because they have $>90\%$ identity to another sequence in the alignment.

[illegible]

[illegible]

[illegible]

Avi01 CCUAAAGUGGAGGCGUAUCCGGU---CAA---ACGNCACCGAUG---AAA**AGGAAU**AGGUGAGACGAGCG---AAGUAA---GGGCGGAAA
Bfr01 UUAUCUUGUCUCCUUAUCCGU---UCAUCCU---CUAUCGGAUU---AAA**AGGAAU**AGGUGU---AAU
Bmg01 AUCAAACAGCAACAGUAAAGGU---GCG---AGUU---AAU**AGGAAU**AGGUGA---AAGA
Lma01---UCAC---AGGUG---AA**AGGAAU**AGGUGAGAGUUGUCCUUAUCUCAAAGGCGGUGAGGAGU---AAU
Pfr01 UGUGUAGGCUAGUAGUCCUUGGU---CCAC---GGCAGUGGUCGCAAG**AGGAAU**AGGUGU---AAU
Rca01 GCCACUCAGGCGGCGGCGUGGU---CUAU---GACAGGCG---AA**AGGAAU**AGGUGAAGGAAUUGGAGCGGUUU---GGGCGGAACCCGA
Rca02 GGCUAUCCAAACAGGCGAUCCG---AACUGGAC---GGAUU---AAU**AGGAAU**AGGUGAGAGAUUAAGG---AUC---GGGCGGAUU
Rca03 GGGCACCUUCGCGGCAUUGGU---CAC---GCGCGCGCGGUG---AA**AGGAAU**AGGUGAGGUGUAAGG---AUC---GGGCGGAAAAU
Rsp01 UGUUUUGGCAAGGCGUACGNG---GCGCG---GGGAG---AAU**AGGAAU**AGGUGG---AAU
Sbi01 UAGACUGCGCCCAUCCAGGUG---CAUG---GGCGAGGUG---AA**AGGAAU**AGGUGAGCGUG---AUC---AGGCGCAGG
Sgr01 UAGCUGAGCGGUGGAGGUGGU---GCCGUGUGG---GCGAGGUGUGCGCAAG**AGGAAU**AGGUGG---AAU
Svi01 AUCGCGCGCUUACGGAACCGU---GAGA---GGGUGAGU---AA**AGGAAU**AGGUGAGGGA---UUAAU---UGGUGAUG
Zmo01 CGGAAUUUUUUUGCAUAGGGG---GAGU---GAGGAA---AAU**AGGAAU**AGGUGG---AAAA
Zmo02 AGCAAUGAGGAAGGAUUAAGGU---CAUUG---GAAAGGU---AA**AGGAAU**AGGUGGGA---GAU---UUUCAAAG

SS	CC	RC	Gy	CCCC	RC	GURa	R
Atu01	CCGGA	ACUG	CCCC	CCGGA	CCUG	UGA	CCGG
Atu02	CCUG	CCUG	CCCC	CCGGA	CCUG	UAA	CCGG
Atu03	CCUG	CCUG	CCCC	CCGGA	CCUG	UGA	CCGG
Atu04	CCGGA	ACUG	CCCC	CCGGA	CCUG	UAA	CCGG
Atu05	CCGGA	ACUG	CCCC	CCGGA	CCUG	UAA	CCGG
Atu06	CCGGA	ACUG	CCCC	CCGGA	CCUG	UAA	CCGG
Bha01	CCGGA	ACUG	CCCC	CCGGA	CCUG	UAA	CCGG
Bha02	CCGGA	ACUG	CCCC	CCGGA	CCUG	UAA	CCGG
Bha03	CCGGA	ACUG	CCCC	CCGGA	CCUG	UAA	CCGG
Bha04	CCGGA	ACUG	CCCC	CCGGA	CCUG	UAA	CCGG
Bsu01	CCGGA	ACUG	CCCC	CCGGA	CCUG	UAA	CCGG
Bja01	CCGGA	ACUG	CCCC	CCGGA	CCUG	UAA	CCGG
Bja02	CCGGA	ACUG	CCCC	CCGGA	CCUG	UAA	CCGG
Bja03	CCGGA	ACUG	CCCC	CCGGA	CCUG	UAA	CCGG
Bja04	CCGGA	ACUG	CCCC	CCGGA	CCUG	UAA	CCGG
Bja05	CCGGA	ACUG	CCCC	CCGGA	CCUG	UAA	CCGG
Bme01	CCGGA	ACUG	CCCC	CCGGA	CCUG	UAA	CCGG
Bme02	CCGGA	ACUG	CCCC	CCGGA	CCUG	UAA	CCGG
Bme03	CCGGA	ACUG	CCCC	CCGGA	CCUG	UAA	CCGG
Bme04	CCGGA	ACUG	CCCC	CCGGA	CCUG	UAA	CCGG
Cer01	CCGGA	ACUG	CCCC	CCGGA	CCUG	UAA	CCGG
Cer02	CCGGA	ACUG	CCCC	CCGGA	CCUG	UAA	CCGG
Cie01	CCGGA	ACUG	CCCC	CCGGA	CCUG	UAA	CCGG
Cie02	CCGGA	ACUG	CCCC	CCGGA	CCUG	UAA	CCGG
Cie03	CCGGA	ACUG	CCCC	CCGGA	CCUG	UAA	CCGG
Cie04	CCGGA	ACUG	CCCC	CCGGA	CCUG	UAA	CCGG
Cie05	CCGGA	ACUG	CCCC	CCGGA	CCUG	UAA	CCGG

Cae01 **CGUGAGCGGU** **CCGCGCA** **CUGUAAU** **AAAGG** **AGUUU** **AGUAGAAU** **AAU**
 Cae02 **CCAAU** **GCAG** **CCGCGU** **UAGU** **UAU** **ACAGUU** **ACAAAGCA** **AG**
 Cpe01 **CCCA** **CGCGU** **CCGCGC** **CUGU** **AA** **UAG** **AGAGG** **UUUU** **UGU** **ACUUUA**
 Cpe02 **CUUUA** **ACUA** **CCUACA** **AA** **CCGUGAA** **GC** **AGACAA** **ACU** **CAAAU**
 Cpe03 **CCAGU** **ACAG** **CCCGCGU** **AGU** **UGU** **AGU** **AGU** **UUA** **AAAG**
 Cpe04 **AAU** **UGA** **CGCGCGCGU** **UA** **UGU** **UUA** **AAAG** **GACGA** **AAU** **AAAG**
 Eco01 **CUGGA** **GCUGA** **CCGCGC** **AGU** **AGU** **UUA** **AAAG** **AAAGU** **CGA** **UUG** **UUA** **AAAG**
 Fru01 **CCAA** **ACAG** **CCGCGC** **AGU** **UUA** **AAAG** **GUUU** **UGU** **UUA** **AAAG**
 Ligo1 **CCGGA** **GCUG** **ACCGCGC** **AGU** **UUA** **AAAG** **AUGAG** **UUUG** **UUA** **AAAG**
 Lmo01 **CCGA** **ACUG** **CCCGCGC** **AGU** **UUA** **AAAG** **GGAA** **AAU** **AAAG**
 Mlo01 **CCGUG** **GCUG** **CCCGCGC** **AGU** **UUA** **AAAG** **UAG** **UUU** **UUA** **AAAG**
 Mlo02 **CCGCG** **GCUG** **CCCGCGC** **AGU** **UUA** **AAAG** **AAAG** **UUU** **UUA** **AAAG**
 Mlo03 **CCGCG** **GCUG** **CCCGCGC** **AGU** **UUA** **AAAG** **UUC** **AAU** **UUA** **AAAG**
 Mlo04 **CCGCG** **GCUG** **CCCGCGC** **AGU** **UUA** **AAAG** **CGA** **UUU** **UUA** **AAAG**
 Mlo05 **CCGUG** **GCUG** **CCCGCGC** **AGU** **UUA** **AAAG** **GAC** **UUU** **UUA** **AAAG**
 Mlo06 **CCGUG** **GCUG** **CCCGCGC** **AGU** **UUA** **AAAG** **AG** **UUU** **UUA** **AAAG**
 Mle01 **CCGGA** **GCUG** **CCCGCGC** **AGU** **UUA** **AAAG** **ACG** **UUU** **UUA** **AAAG**
 Mtu01 **CCGCG** **GCUG** **CCCGCGC** **AGU** **UUA** **AAAG** **GAG** **UUU** **UUA** **AAAG**
 Mtu02 **CCGGA** **GCUG** **CCCGCGC** **AGU** **UUA** **AAAG** **ACG** **UUU** **UUA** **AAAG**
 Pae01 **CCAGC** **GCUG** **CCCGCGC** **AGU** **UUA** **AAAG** **AUC** **UUU** **UUA** **AAAG**
 Pae02 **CCGUG** **GCUG** **CCCGCGC** **AGU** **UUA** **AAAG** **UAU** **UUU** **UUA** **AAAG**
 Pae03 **CCGCG** **GCUG** **CCCGCGC** **AGU** **UUA** **AAAG** **CGA** **UUU** **UUA** **AAAG**

[illegible]

Avi01 CCGUGGCGUGCCCCGCAACUGUGA-ACGG- CGAGCGAUGGCGAUCAU-----
Bfr01 CCGGGACAGU-CCCGCGUGGUGUGAA-CCUCC- GUCUGAAUUUGCGAGUAGCAACUCGUG-
Bmg01 CCAGUACUGGCCCCGCAACUGUAA-GUGUG- GACGAAGGAGUUA-----
Lma01 CCGGUGGUGCCCCCGCAACCGUAA-CCGAG- UUAAGCGUGGAAU-----
Pfr01 CCGGAACUGU-CCCGGAGGUGCA-AUGGG- AACGACGAAAGUAAAG-----
Rsa01 CCGGAACUGGCCCCCGCAACCGUGACCGGA- GAGGCGCCCGGAG-----
Rsa02 CCGUGGCGUGCCCCCGCAACUGUGA-CCGG- CGAGACGAGCGGUGCGAAG-----
Rsa03 CCGUACUGGCCCCCGCAACUGUAA-CCGG- CGAGGAGGCGCGCCCGCA-----
-Rsp01 CCGGGCGGCG-CCCGCGGCGUGUGA-CCG- GGAUGGCGCGGGGAAGAG-----
Sbi01 CCGGGCGGCGCCCCCGCAACGGUAA-CCACGUC- AGUCGAGGGAACAAC-----
Sgi01 CCGGGACUGG-CCCGGCAACGGUGA-GUGGG- AACGACCGCGGUGCAUA-----
Shi01 CCGAGACUGGCCCCCGCAACUGUAA-CCGG- AGAGUGAUGGUGGUAUCUACGGAUUA-----
Zma01 CCGUGGCGGCCCCGCAACUGUAA-ACAGU- UGAAAGGCGAATA-----
Zma02 CCGUGGCGGCCCCCGCAACUGUAA-ACGG- CGAGCAAGGATCAAAU-----

SS	YGGGAAGGY
Cons	tyCACUG
Auu01	CAAAA
Auu02	UUCG
Auu03	CCGUCAUCA
Auu04	CCGG
Auu05	GCAA
Auu06	UUUAUGAUC
Bha01	UUCAU
Bha02	UACUUCU
Bha03	AAA
Bha04	UAAC
Bsu01	CAAGC
Bja01	UGCAC
Bja02	UGAC
Bja03	GCAA
Bja04	UCCUGGU
Bja05	AUCUCGG
Bme01	AAUUA
Bme02	AGAUG
Bme03	CAC
Bme04	GAAA
Ccr01	GCUGG
Ccr02	CAAAA
Cte01	GUUCA
Cte02	UCCAC
Cte03	GCCC
Cte04	AUAAC
Cte05	UCCAC

Strain	Genotype	Phenotype
Cac01	UUA UUG	GAAA
Cac02	UUA UUG	UUUU
Cpe01	UUA UUG	UUUU
Cpe02	UUA UUG	UUUU
Cpe03	UUA UUG	UUUU
Cpe04	UUA UUG	UUUU
Eco01	UUA UUG	UUUU
Fnu01	UUA UUG	UUUU
Lig01	UUA UUG	UUUU
Lmo01	UUA UUG	UUUU
Mib01	UUA UUG	UUUU
Mib02	UUA UUG	UUUU
Mib03	UUA UUG	UUUU
Mib04	UUA UUG	UUUU
Mib05	UUA UUG	UUUU
Mib06	UUA UUG	UUUU
Mib07	UUA UUG	UUUU
Mib08	UUA UUG	UUUU
Mib09	UUA UUG	UUUU
Mib10	UUA UUG	UUUU
Mib11	UUA UUG	UUUU
Mib12	UUA UUG	UUUU
Mib13	UUA UUG	UUUU
Mib14	UUA UUG	UUUU
Mib15	UUA UUG	UUUU
Mib16	UUA UUG	UUUU
Mib17	UUA UUG	UUUU
Mib18	UUA UUG	UUUU
Mib19	UUA UUG	UUUU
Mib20	UUA UUG	UUUU
Mib21	UUA UUG	UUUU
Mib22	UUA UUG	UUUU
Mib23	UUA UUG	UUUU
Mib24	UUA UUG	UUUU
Mib25	UUA UUG	UUUU
Mib26	UUA UUG	UUUU
Mib27	UUA UUG	UUUU
Mib28	UUA UUG	UUUU
Mib29	UUA UUG	UUUU
Mib30	UUA UUG	UUUU
Mib31	UUA UUG	UUUU
Mib32	UUA UUG	UUUU
Mib33	UUA UUG	UUUU
Mib34	UUA UUG	UUUU
Mib35	UUA UUG	UUUU
Mib36	UUA UUG	UUUU
Mib37	UUA UUG	UUUU
Mib38	UUA UUG	UUUU
Mib39	UUA UUG	UUUU
Mib40	UUA UUG	UUUU
Mib41	UUA UUG	UUUU
Mib42	UUA UUG	UUUU
Mib43	UUA UUG	UUUU
Mib44	UUA UUG	UUUU
Mib45	UUA UUG	UUUU
Mib46	UUA UUG	UUUU
Mib47	UUA UUG	UUUU
Mib48	UUA UUG	UUUU
Mib49	UUA UUG	UUUU
Mib50	UUA UUG	UUUU
Mib51	UUA UUG	UUUU
Mib52	UUA UUG	UUUU
Mib53	UUA UUG	UUUU
Mib54	UUA UUG	UUUU
Mib55	UUA UUG	UUUU
Mib56	UUA UUG	UUUU
Mib57	UUA UUG	UUUU
Mib58	UUA UUG	UUUU
Mib59	UUA UUG	UUUU
Mib60	UUA UUG	UUUU
Mib61	UUA UUG	UUUU
Mib62	UUA UUG	UUUU
Mib63	UUA UUG	UUUU
Mib64	UUA UUG	UUUU
Mib65	UUA UUG	UUUU
Mib66	UUA UUG	UUUU
Mib67	UUA UUG	UUUU
Mib68	UUA UUG	UUUU
Mib69	UUA UUG	UUUU
Mib70	UUA UUG	UUUU
Mib71	UUA UUG	UUUU
Mib72	UUA UUG	UUUU
Mib73	UUA UUG	UUUU
Mib74	UUA UUG	UUUU
Mib75	UUA UUG	UUUU
Mib76	UUA UUG	UUUU
Mib77	UUA UUG	UUUU
Mib78	UUA UUG	UUUU
Mib79	UUA UUG	UUUU
Mib80	UUA UUG	UUUU
Mib81	UUA UUG	UUUU
Mib82	UUA UUG	UUUU
Mib83	UUA UUG	UUUU
Mib84	UUA UUG	UUUU
Mib85	UUA UUG	UUUU
Mib86	UUA UUG	UUUU
Mib87	UUA UUG	UUUU
Mib88	UUA UUG	UUUU
Mib89	UUA UUG	UUUU
Mib90	UUA UUG	UUUU
Mib91	UUA UUG	UUUU
Mib92	UUA UUG	UUUU
Mib93	UUA UUG	UUUU
Mib94	UUA UUG	UUUU
Mib95	UUA UUG	UUUU
Mib96	UUA UUG	UUUU
Mib97	UUA UUG	UUUU
Mib98	UUA UUG	UUUU
Mib99	UUA UUG	UUUU
Mib100	UUA UUG	UUUU
Mib101	UUA UUG	UUUU
Mib102	UUA UUG	UUUU
Mib103	UUA UUG	UUUU
Mib104	UUA UUG	UUUU
Mib105	UUA UUG	UUUU
Mib106	UUA UUG	UUUU
Mib107	UUA UUG	UUUU
Mib108	UUA UUG	UUUU
Mib109	UUA UUG	UUUU</

[illegible]

Aca01	GGGAGGAGGAG	GCAUAG	UUUUUUU GAA UUUU
Avi01	GGGAGGAGGAG	CCAC	UUUUUUU UAA UUU
Bfr01	GGGAGGAGGAG	UUAA	UUUUUUU GAA UUUU
Bmg01	GGGAGGAGGAG	AAAA	UUUUUUU GAA UUUU
Lma01	GGGAGGAGGAG	UCCAG	UUUUUUU GAA UUU
Pfr01	GGGAGGAGGAG	GCAA	UUUUUUU GAA UUU
Rca01	GGGAGGAGGAG	ACCA	UUUUUUU GAA UUU
Rca02	GGGAGGAGGAG	AUCCAC	UUUUUUU GAA UUU
Rca03	GGGAGGAGGAG	ACCG	UUUUUUU GAA UUU
Rsp01	GGGAGGAGGAG	UUC	UUUUUUU GAA UUU
Sbi01	GGGAGGAGGAG	ACG	UUUUUUU GAA UUU
Sgi01	GGGAGGAGGAG	GGAC	UUUUUUU GAA UUU
Svi01	GGGAGGAGGAG	UUCG	UUUUUUU GAA UUU
Zmc01	GGGAGGAGGAG	UCUA	UUUUUUU GAA UUU
Zmc02	GGGAGGAGGAG	UUAU	UUUUUUU GAA UUU

SS	Cons	yy RAGYC GRAgACC GCC
Aiu01	GAUAAGAUUUGUGAC	CGGC-AGGUCAGGAGACCUGCCUUGAGCGGAAAUGUCCACG
Aiu02	GAUGAGGCACGCA	AUUGGU-CGGCGAGGAGACCUGCCGCGGAAAUUGGAAACCAUCCG
Aiu03	UGCCGGGAGGUGUUCAGGUUUUGAC	CCGU-AGGCGAGGAGACCUGCCGCGGAAAGAUUACGUCCACG
Aiu04	GGACCAAGGUAUGACC	CGC-AGGCGAGGAGACCUGCCGCGGAAAGAUUACGUCCACG
Aiu05	GAAAGGGGCGAUGAA	CCU-UUGUCAGAAAGACCGGCGGCGGAAAGAUUACGUCCACG
Aiu06	GGUGACAGGUGUUGAU	CCCGG-AGGCGAGGAGACCUGCCGCGGAAAGAUUACGUCCACG
Bha01	UGUUGCAUUGUGAC	CGUA-AGGCGAGGAGACCUGCCGCGGAAAGAUUACGUCCACG
Bha02	UUCUAAAGUAGGUA	AGCAG-CAAGCGAGGAGACCUGCCGCGGAAAGAUUACGUCCACG
Bha03	CACAUGGAGUGUGA	UCUU-AGGCGAGGAGACCUGCCGCGGAAAGAUUACGUCCACG
Bha04	CCAAGUAUCUGAUGAA	CAU-AGGCGAGGAGACCUGCCGCGGAAAGAUUACGUCCACG
Bsu01		GCCAA-GAGCAGGAUAACCUGCCGCGGAAAGAUUACGUCCACG
Bja01	GGGACAGGGAUGA	CCAGC-AGGCGAGGAGACCUGCCGCGGAAAGAUUACGUCCACG
Bja02	GGGGAUUGAAGGCAAAACCCUG	CUCCG-CAAGCGGAGAGACCUGCCGCGGAAAGAUUACGUCCACG
Bja03	CGGAGAAAGCCAGUCA	AGCGC-GAGCAGGAGAGACCUGCCGCGGAAAGAUUACGUCCACG
Bja04	GACGAAUGUGUGACGAC	CCGC-GAGCAGGAGAGACCUGCCGCGGAAAGAUUACGUCCACG
Bja05	GACGUAAGGUACGAC	CCGC-GAGCAGGAGAGACCUGCCGCGGAAAGAUUACGUCCACG
Bme01	GGAGGCAAGGCAAGAC	CCCG-AGGCGAGGAGACCUGCCGCGGAAAGAUUACGUCCACG
Bme02	GGUUAUUGCCCAAGAC	CCGC-CAAGCGGAGAGACCUGCCGCGGAAAGAUUACGUCCACG
Bme03	GAUGGACGGGAUA	UCGCG-AGGCGAGGAGAGACCUGCCGCGGAAAGAUUACGUCCACG
Bme04	GGGUGGAAGGUGUA	CCCG-AGGCGAGGAGAGACCUGCCGCGGAAAGAUUACGUCCACG
Cer01	GGGCAAGGUGACGAC	CCGU-AGGCGAGGAGAGACCUGCCGCGGAAAGAUUACGUCCACG
Cer02	GACGCCCCAGAGCAUUGAC	CCGU-AGGCGAGGAGAGACCUGCCGCGGAAAGAUUACGUCCACG
Cle01	CGGCAGAUUC	CGGA-AGGCGAGGAGAGACCUGCCGCGGAAAGAUUACGUCCACG
Cle02	UGGAGGGGCGUGGUA	CCCGGUA-AGGCGAGGAGAGACCUGCCGCGGAAAGAUUACGUCCACG
Cle03	GGAAAGCGA	CCGGA-AGGCGAGGAGAGACCUGCCGCGGAAAGAUUACGUCCACG
Cle04	UGCCCCGAGGAAACGUGCAAGUAUUGCA	CCCAUC-GAGGCGAGGAGAGACCUGCCGCGGAAAGAUUACGUCCACG
Cle05	UGCAUUG	CCGU-AAAGCGAGGAGAGACCUGCCGCGGAAAGAUUACGUCCACG

Cae001 GUACUUAAGCA AUGA
Cae002 UGGUUGAGGCCUA
Cpe001 ACAAAAGUGAUGAUA
Cpe002 GAGAUUGAGGAAGAA
Cpe003 AGAAAUAGAGGAUAAG
Cpe004 UUGUUGUAGGAUGA
Eae001 AUCUUCUUAAGUUCUUAAGAGCCC
Fnu001 AAGAAUAUAAA
Lig01 UGCGAAA
Lmo01 CGAUUUGUUGAUGAA
Mlo01 GGGGAAGGGGCCUGAU
Mlo02 CGGCACCAAGGGCGAUGACC
Mlo03 GUCGCGACCGUCCGCAAGGAGA
Mlo04 GGGCAGAGGCUUUGAC
Mlo05 ACCGGACGGGGUUGA
Mlo06 GGAUUGGCUAGAGC
Mlo07 AUGGCCAUUAGAAGCAUUAUCCAGUGGCG
Mtu01 GAGGGCAAGCAACGA
Mtu02 GAGCGGCCAGUAAGGAGGAGCCGACCGGGUUGGAG
Pae01 GCGGCUUGAAAGCGUCCA GCGCUUGC
Pae02 GCGCGCAAGCGGAGGUUCUCCCGCGGUGGAGAGC
Pae03 GAGGAUUGACGAGC
Cae001 GUACUUAAGCA AUGA
Cae002 UGGUUGAGGCCUA
Cpe001 ACAAAAGUGAUGAUA
Cpe002 GAGAUUGAGGAAGAA
Cpe003 AGAAAUAGAGGAUAAG
Cpe004 UUGUUGUAGGAUGA
Eae001 AUCUUCUUAAGUUCUUAAGAGCCC
Fnu001 AAGAAUAUAAA
Lig01 UGCGAAA
Lmo01 CGAUUUGUUGAUGAA
Mlo01 GGGGAAGGGGCCUGAU
Mlo02 CGGCACCAAGGGCGAUGACC
Mlo03 GUCGCGACCGUCCGCAAGGAGA
Mlo04 GGGCAGAGGCUUUGAC
Mlo05 ACCGGACGGGGUUGA
Mlo06 GGAUUGGCUAGAGC
Mlo07 AUGGCCAUUAGAAGCAUUAUCCAGUGGCG
Mtu01 GAGGGCAAGCAACGA
Mtu02 GAGCGGCCAGUAAGGAGGAGCCGACCGGGUUGGAG
Pae01 GCGGCUUGAAAGCGUCCA GCGCUUGC
Pae02 GCGCGCAAGCGGAGGUUCUCCCGCGGUGGAGAGC
Pae03 GAGGAUUGACGAGC

[illegible]

Avf01 GGAACAAAGCCAGAC CCGU GAGCCAGGAGAG CCGUCCCGAUAAAGCAUGCCGGAAGC
 Bf01 GUCCGAAACAA GGAGU CAGUCAGAAAGA CCGUCCCGCUUAUCAAAAGGCUGUUUC
 Bmg01 CUCAAAGUAGAAUGA UACAG AAGUCAGGAGAG CCGUCCCGAUUAUUGUGAAGUUUCUUAU
 Lma01 UGAUGCUUUAAGAAUGA CUCGC AAGCCCGGAGAG CCGCCCGGA AAAAUCAGAUAAACA
 Phf01 GUAGUGGAGGAAGUCCGGAGUGAUUUGGGAUG CCGAU GAGUCCGAAGA CCGUCCCGAGCGGACCAUCUGU
 Rca01 GGGGGGACCGUUGAGGGGACCCCGCUUGCA UCCG CAAAGCCCGGAGAG CCGUCCCGAGCGGAGUGGAUUCGGGCG
 Rca02 GACCGGAGUUGAUUCGAA CCGC AAGUCAGGAGAG CCGUCCCGAUCCGCUUGGCGUGGCAAG
 Rca03 GGGGGAAGGCAAGAC CCGC AAGUCAGGAGAG CCGUCCCGAUCCGCUUGGCGUGGCAAG
 Rsp01 GCGGCGGCGGAGUAG CCG GAGCGAGAGAG CCGCGCGGUGAGGCGAGGUGUCCCGCC
 Sbi01 GCGGCGGCGGAGUAG CCG GAGCGAGAGAG CCGCGCGGUGAGGCGAGGUGUCCCGCC
 Sgi01 GCGGCGGCGGAGUAG CCG GAGCGAGAGAG CCGCGCGGUGAGGCGAGGUGUCCCGCC
 Svi01 GCGGCGGCGGAGUAG CCG GAGCGAGAGAG CCGCGCGGUGAGGCGAGGUGUCCCGCC
 Zma01 GGUUGUUGAU CCGU CAGGAGAG CCGUCCCGAUUAAGCAAGUCAUCCACU
 Zma02 UGAUCCGACCGCGUGAC CCGU CAGGAGAG CCGUCCCGAUUAAGCAAGUCAUCCACU

C. G-Box

NC_002570.1/648448-648540 *Bacillus halodurans*
 NC_002570.1/650317-650406 *Bacillus halodurans*
 NC_002570.1/676483-676572 *Bacillus halodurans*
 NC_002570.1/806882-806965 *Bacillus halodurans*
 NC_002570.1/1593067-1592976 *Bacillus halodurans*
 NC_000964.1/693955-694038 *Bacillus subtilis*
 NC_000964.1/697886-697976 *Bacillus subtilis*
 NC_000964.1/2319120-2319031 *Bacillus subtilis*
 NC_000964.1/4004319-4004410 *Bacillus subtilis*
 NC_003030.1/1002184-1002270 *Clostridium acetobutylicum*
 NC_003030.1/2904259-2904168 *Clostridium acetobutylicum*
 NC_003030.1/2824539-2824454 *Clostridium acetobutylicum*
 NC_003366.1/512410-512323 *Clostridium perfringens*
 NC_003366.1/7617892-2617807 *Clostridium perfringens*
 NC_003454.1/1645257-1645173 *Fusobacterium nucleatum*
 NC_002662.1/1159519-1159604 *Lactococcus lactis* subsp. *lactis*
 NC_003210.1/630773-630678 *Listeria monocytogenes* strain EGD
 NC_003210.1/1595601-1958511 *Listeria monocytogenes* strain EGD
 NC_004193.1/760480-760571 *Oceanobacillus thelyensis*
 NC_004193.1/769695-769781 *Oceanobacillus thelyensis*
 NC_004193.1/786775-786863 *Oceanobacillus thelyensis*
 NC_004193.1/1103947-1104044 *Oceanobacillus thelyensis*
 NC_002745.1/430771-430861 *Staphylococcus aureus* subsp. *aureus*
 NC_004461.1/2433384-2432294 *Staphylococcus epidermidis*
 NC_004116.1/1093950-1093860 *Streptococcus agalactiae*
 NC_002737.1/930757-930842 *Streptococcus pyogenes*
 NC_003028.1/1754793-1754878 *Streptococcus pneumoniae*
 NC_003669.1/5961672-5961663 *Thermanserobacter tengcongensis*

Consensus

[illegible]

D. A-Box

NC_000564.1/626134-626051 *Bacillus subtilis*
NC_003366.1/2870819-2870732 *Clostridium perfringens*
NC_004460.1/504378-504467 *Vibrio vulnificus*

AATTAAATAGCTATATCATCTGGATTAACCTCAATTAATATCG..TTCGCGTTCCTACCTGGAA..CCGTAAATCCCG..ATTCGCAATTTGTTATGACATTT
ATPAAAAATTAATTTTGGCTTCTATTAACCTCAATTAATATCG..TTCGCGTTCCTACCTGGAA..CCGTAAATCCCG..ATTCGCAATTTGTTATGACATTT
GACTTCGGCGATCAACGCTTCAATTAACCTCAATTAATATCG..TTCGCGTTCCTACCTGGAA..CCGTAAATCCCG..ATTCGCAATTTGTTATGACATTT

E. Lysine riboswitch comparison

Command-1 Plain Text

Command-2 Base paired stem 1

Command-3 Base paired stem 2

i. Command-4 Base paired stem 3

Command-5 Base paired stem 4

Command-6 Base paired stem 5

ii. Command-7 Base paired stem 6

Command-8 Terminator poly-U

Command-9 Downstream A1 stem paired to stem 1

Command-10 Optional base paired stem 2

CUAG is 90% sequence similarity

2. CUAG is the Anti-Terminator

CUAG is the Terminator stem

1. Bha LysC AGUGAUGGUA AGGU **AKGAAACG** AAG-AGUAC **ACAGUGUGAGAGAUAUG** AGAAU **CCUUGAC** **GACUGUUGGAAAGG** **GGGAUUG** **CGAGUGGUAAGUICGGG** CUCAU **UGCG** **AUUUGCGG** **UAACUAGUU** **GAAUA-AGCAUAGGCG** **UACACAGACUAU** **CCCCAA** **UAGUGCGUGUGAGAA** **UAUUCACAGU**
2. Bha dda AGUGAGAGUA AGGU **FAAAACG** AAG-AGUA **CACAUUGGA** **GAU** **GAAUGAGA** **UUGUGAGACAUUGUG** **GAAAGG** **GAUUG** **CGAGUGGUAAGUICGGG** CUCAU **UGCG** **AUUUGCGG** **UAACUAGUU** **GAAUA-AGCAUAGGCG** **UACACAGACUAU** **CCCCAA** **UAGUGCGUGUGAGAA** **UAUUCACAGU**
3. Bha hnc AGUGAGAGUA AGGU **FAAGGUUU** AAG-AGUA **AGCGUUU** **GAGUAGACACGAGGA** **UAAGCG** **CGAAAGG** **AAACUCG** **CGAGUGGUAAGUICGGG** CUCAU **UGCG** **AUUUGCGG** **UAACUAGUU** **GAAUA-AGCAUAGGCG** **UACACAGACUAU** **CCCCAA** **UAGUGCGUGUGAGAA** **UAUUCACAGU**
4. Bha LysC GGUAGAGUAU AGGU **FGUA** **GUU** **UUG** **UAG** **AG** **AGUA** **AGCGUUU** **GAGUAGACACGAGGA** **UAAGCG** **CGAAAGG** **AAACUCG** **CGAGUGGUAAGUICGGG** CUCAU **UGCG** **AUUUGCGG** **UAACUAGUU** **GAAUA-AGCAUAGGCG** **UACACAGACUAU** **CCCCAA** **UAGUGCGUGUGAGAA** **UAUUCACAGU**
5. Cpe lya ACCUUGUGUA AGU **FGUUAAGU** AAG-AGUA **AGCGUUU** **GAGUAGACACGAGGA** **UAAGCG** **CGAAAGG** **AAACUCG** **CGAGUGGUAAGUICGGG** CUCAU **UGCG** **AUUUGCGG** **UAACUAGUU** **GAAUA-AGCAUAGGCG** **UACACAGACUAU** **CCCCAA** **UAGUGCGUGUGAGAA** **UAUUCACAGU**
6. Cpe hnc AGAGAGUA AGG **FGAGAUUG** AAG-AGUA **AGCGUUU** **GAGUAGACACGAGGA** **UAAGCG** **CGAAAGG** **AAACUCG** **CGAGUGGUAAGUICGGG** CUCAU **UGCG** **AUUUGCGG** **UAACUAGUU** **GAAUA-AGCAUAGGCG** **UACACAGACUAU** **CCCCAA** **UAGUGCGUGUGAGAA** **UAUUCACAGU**
7. Cpe lya GACCAAGUA AGU **FGUAUUAU** AAG-AGUA **AGCGUUU** **GAGUAGACACGAGGA** **UAAGCG** **CGAAAGG** **AAACUCG** **CGAGUGGUAAGUICGGG** CUCAU **UGCG** **AUUUGCGG** **UAACUAGUU** **GAAUA-AGCAUAGGCG** **UACACAGACUAU** **CCCCAA** **UAGUGCGUGUGAGAA** **UAUUCACAGU**
8. Cpe lya AACUGAGUA AGG **FGUAUUAU** AAG-AGUA **AGCGUUU** **GAGUAGACACGAGGA** **UAAGCG** **CGAAAGG** **AAACUCG** **CGAGUGGUAAGUICGGG** CUCAU **UGCG** **AUUUGCGG** **UAACUAGUU** **GAAUA-AGCAUAGGCG** **UACACAGACUAU** **CCCCAA** **UAGUGCGUGUGAGAA** **UAUUCACAGU**
9. Bha hnc CAGGAGUA AGG **FGUAUUAU** AAG-AGUA **AGCGUUU** **GAGUAGACACGAGGA** **UAAGCG** **CGAAAGG** **AAACUCG** **CGAGUGGUAAGUICGGG** CUCAU **UGCG** **AUUUGCGG** **UAACUAGUU** **GAAUA-AGCAUAGGCG** **UACACAGACUAU** **CCCCAA** **UAGUGCGUGUGAGAA** **UAUUCACAGU**
10. Bha hnc UAGAGUA AGU **FGUAUUAU** AAG-AGUA **AGCGUUU** **GAGUAGACACGAGGA** **UAAGCG** **CGAAAGG** **AAACUCG** **CGAGUGGUAAGUICGGG** CUCAU **UGCG** **AUUUGCGG** **UAACUAGUU** **GAAUA-AGCAUAGGCG** **UACACAGACUAU** **CCCCAA** **UAGUGCGUGUGAGAA** **UAUUCACAGU**
11. Bha dda GGUUGUAU AGU **FGUAUUAU** AAG-AGUA **AGCGUUU** **GAGUAGACACGAGGA** **UAAGCG** **CGAAAGG** **AAACUCG** **CGAGUGGUAAGUICGGG** CUCAU **UGCG** **AUUUGCGG** **UAACUAGUU** **GAAUA-AGCAUAGGCG** **UACACAGACUAU** **CCCCAA** **UAGUGCGUGUGAGAA** **UAUUCACAGU**
12. Bha hnc UCGUGUGUA AGG **FGUAUUAU** AAG-AGUA **AGCGUUU** **GAGUAGACACGAGGA** **UAAGCG** **CGAAAGG** **AAACUCG** **CGAGUGGUAAGUICGGG** CUCAU **UGCG** **AUUUGCGG** **UAACUAGUU** **GAAUA-AGCAUAGGCG** **UACACAGACUAU** **CCCCAA** **UAGUGCGUGUGAGAA** **UAUUCACAGU**
13. Pms hnc UACUGUAU AGG **FGUAUUAU** AAG-AGUA **AGCGUUU** **GAGUAGACACGAGGA** **UAAGCG** **CGAAAGG** **AAACUCG** **CGAGUGGUAAGUICGGG** CUCAU **UGCG** **AUUUGCGG** **UAACUAGUU** **GAAUA-AGCAUAGGCG** **UACACAGACUAU** **CCCCAA** **UAGUGCGUGUGAGAA** **UAUUCACAGU**

[illegible]